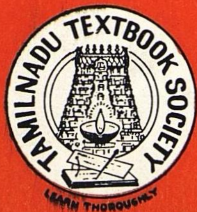
STATISTICS

HIGHER SECONDARY
SECOND YEAR



TAMILNADU
TEXTBOOK SOCIETY

STATISTICS

Higher Secondary—Second Year



TAMILNADU TEXTBOOK SOCIETY
MADRAS

First Edition—1981

Editorial Board Chairman
(Author & Review Committee Member)

Thiru. **M. Sankaranarayanan**, M.A., B.Sc.,
Joint Director of Statistics,
Department of Statistics,
MADRAS-600 006.

Review Committee Members:

Thiru. **T. K. Manickavachagam Pillai**, M.A., L.T.,
Professor of Mathematics (Retd.)
A.C. College of Technology,
MADRAS-600 035.

Thiru. **R. Hanumantha Rao**, M.A.,
Professor of Mathematics,
P.S.G. Arts College,
COIMBATORE.

Price : Rs. 9 - 00

This book has been printed on concessional paper of 60 G.S.M. substance made available by the Government of India.

Printed at
MANI PRINTERS, MADRAS-600 010.

CONTENTS

FIRST PAPER

SECOND PAPER

II YEAR — 1st PAPER

CHAPTER I

PROBABILITY

The term 'Probability' means likelihood or chance or possibility. It can be used qualitatively and quantitatively. However, in statistics this term is used in quantitative sense only. It is advantageous at this stage to be familiar with certain terms which are generally used in the study of probabilities.

Experiment or Trial

Tossing of a coin or die is generally meant as an experiment or trial. A tossing of a single coin or die for 5 times means 5 experiments or 5 trials. It is also equivalent to tossing five coins or five dice at a time.

Examples of Events

A coin has two sides, namely head and tail. When a coin is tossed, any one of the two sides, either head or tail, may turn up and turning up of head or tail is denoted as an event. A die has six sides marked, say 1, 2, 3, 4, 5 and 6. When a die is thrown on a table the upper face may be any one of the series marked 1, 2, 3, 4, 5 and 6. The coming up of the upper face marked either 1 or 2 or 3...6 is considered to be an event.

Exhaustive Events

When a coin is tossed either head or tail may turn up. These two are the only and possible events that can take place. A group of events is said to be exhaustive if it includes all possible events.

Equally Likely Events

When a coin is tossed, we cannot say which side will turn up since either head or tail may turn up. We have no reason

to believe or predict which side will turn up. This means both head and tail have equal chances of turning up. Hence these two events are said to be equally likely events.

Mutually Exclusive events

Two events are said to be mutually exclusive when the occurrence of one event prevents the occurrence of the other event. When a coin is tossed either a head or tail may turn up and on no occasion both the head and tail can turn up simultaneously. Therefore, these two events namely turning up of head and turning up of tail are mutually exclusive events.

Independent events

The events are said to be independent when the occurrence of one event in a trial does not affect the occurrence of the event in the next or succeeding trials.

Compound Events

When two or more events occur simultaneously, it is said to be compound events. If 2 coins are tossed simultaneously, we may get two heads or two tails or one head and one tail.

Definition of Probability

There are two types of probability. They are: (1) Mathematical probability or apriori probability; (2) Statistical probability or Empirical probability or aposteriori probability.

Mathematical probability (or) Apriori probability.

Apriori probability is one which can be determined prior to any experimentation or trial. It is based on the following assumptions. (i) We have full confidence that the event will happen out of several possible alternatives which are *mutually exclusive*. (ii) The various possible alternatives are equally likely. Hence Mathematical probability is equal to the total number of cases favourable for an occurrence of the event divided by the total number of all possible cases.

$$\text{Probability} = \frac{\text{Number of favourable cases}}{\text{Total number of possible cases}}$$

Tossing of coin

Suppose a coin is tossed, there are only two ways either head may turn-up or tail may turn-up. These two events are mutually exclusive and equally likely. Hence the probability of getting head is equal to $\frac{1}{2}$.

Let us find the probability of the occurrence of two heads in a throw of 2 coins.

Each coin contains two faces, one head denoted by letter 'H' and the other, tail denoted by 'T'. In a throw of a coin, any of the two sides either H or T may occur.

So there are four possible ways of occurrence of the heads and tails in a throw of 2 coins.

(1)	(2)	(3)	(4)
H.H.	H.T.	T.T.	T.H.

Out of these 4 ways, only one case is considered to be a success. Hence the probability of the success is $\frac{1}{4}$.

Similarly in a throw of 3 coins the probability of 2 heads and one Tail (HHT) is $\frac{3}{8}$ as explained below:

There are 8 ways for the occurrence of heads or tails.

HHH		
HHT	HTH	THH
TTT		
TTH	THT	HTT

It should be noted that a trial or throw of 2 coins or 3 coins means the simultaneous throw of all the 2 or 3 coins. The throw of one coin for 2 times or three times as the case may be, may also be considered as one throw or one trial.

Throw of Dice

A die is in cubical shape having six sides marked 1,2, 3, 4, 5 & 6. Suppose a die is thrown, let us find out the probability of getting the face marked 4.

Total number of faces in the die $= 6$.

Number of faces marked 4 $= 1$.

The probability of getting the face marked 4 $= \frac{1}{6}$

Two dice are thrown. Find the probability for the sum of the numbers occurring on the faces is equal to 10.

Each die has six sides or faces and the numbers written on the faces are 1, 2, 3, 4. 5 & 6. If one die is thrown once, any one of the six sides may appear. Similarly in the case of other die also any one of the six sides may occur. But, for the appearance of one face in a die there will be six ways or six faces for the second die. But, for the first die alone there are six ways. Therefore, the total number of ways in which the faces of two dice occur or appear simultaneously is 36. Of these 36 ways, only in 3 ways as given below we can get the sum equal to 10.

(4 & 6) (5 & 5), or (6 & 4).

Therefore the probability is $\frac{3}{36} = \frac{1}{12}$

Let us take a throw of 3 dice and find the probability for getting the sum equal to 13.

The face of a die can appear in six ways independently. However, the simultaneous appearance of faces of 2 dice will be equal to $6 \times 6 = 36$ ways. Similarly the simultaneous occurrence of faces in the three dice will be $36 \times 6 = 216$ ways. We can get the sum equal to 13 in the following manner:

(4, 4 & 5)	3, 4, 6	2, 5, 6	6, 6, 1	5, 5, 3
(4, 5 & 4)	3, 6, 4	2, 6, 5	6, 1, 6	5, 3, 5
(5, 4 & 4)	6, 4, 3	5, 2, 6	1, 6, 6	3, 5, 5

6 3 4	5 6 2
4 3 6	6 2 5
4 6 3	6 5 2

$$\text{Probability} = \frac{21}{216} = \frac{7}{72}$$

Mathematical probability is useful in the case of games of chance like tossing of coins or throwing dice etc., where we can find out the total number of equally likely causes without actually conducting the experiment. But in actual life, the possible cases are not equal for any kind of events. Hence the mathematical probability is not suitable in such cases. In the case of chance of events, we cannot find out the number of favourable cases and the total number of possible cases. In such cases the probability is to be determined with the help of facts and figures of past observations only.

Aposteriori (or) Statistical Probability (or) Empirical Probability

Probability based on past experience from a long series of experiments is known as Statistical or Empirical or aposteriori probability.

Suppose we conduct an experiment and repeat the same experiment under the same set of identical conditions for a large number of times say 'n' times. Let us also count the number of times a particular event occurs and let us suppose it to be 'm' times. Then the ratio $\frac{m}{n}$ is called the statistical probability.

If, consistent with a given set of conditions, there are 'n' exhaustive and mutually exclusive and equally likely causes, and 'm' of the causes are favourable for an occurrence of an event 'E' then the probability of the event 'E' is denoted by the symbol P(E) and it is defined as $\frac{m}{n}$

$$\text{ie.}: \quad P(E) = \frac{m}{n}$$

Difference between Mathematical and Statistical Probabilities

1. Mathematical probability is determined without conducting experiments. But statistical probability is determined after conducting the experiments and the results obtained.

2. Mathematical probability can be expressed as an exact correct quantity. But statistical probability is only an estimate and hence it is only an approximation.

Limits of the value of Probability

The probability for occurrence of an event may be denoted by the letter 'p' and the probability for the occurrence of the failure by the letter 'q' such that p + q will be equal to 1.

$$p + q = 1$$

$$p = 1 - q$$

$$q = 1 - p$$

Therefore, probability may be any number between 0 and 1. If the probability is equal to 1 then it denotes the absolute certainity of the event. Similarly, if it is 0, then it denotes the absolute impossibility or the failure of the event.

Calculation of Probability

Though statistical definition of probability is very useful, in practice, we use only the definition of mathematical probability in our calculations.

Combination

Before we proceed further, we shall know something about combination. Let us suppose that there are 5 players and we want to choose only 3 players for our purpose. This can be done in 10 ways. But the answer will be written in a symbolic form as follows :

$5C_3$: It means the number of combinations of 5 items taken 3 at a time.

$$5C_3 = \frac{5 \times 4 \times 3}{1 \times 2 \times 3} = 10 = \frac{5\times4\times3\times2\times1}{1\times2\times3\times1\times2}$$

Similarly, $$7C_3 = \frac{7 \times 6 \times 5}{1 \times 2 \times 3} = 35$$

$$= \frac{7\times6\times5\times4\times3\times2\times1}{1\times2\times3\times1\times2\times3\times4}$$

$$8C_2 = \frac{8 \times 7}{1 \times 2} = 28$$

$$= \frac{8\times7\times6\times5\times4\times3\times2\times1}{1\times2\times1\times2\times3\times4\times5\times6}$$

$$10c_4 = \frac{10 \times 9 \times 8 \times 7}{1 \times 2 \times 3 \times 4} = 210$$

$$= \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{1\times2\times3\times4\times1\times2\times3\times4\times5\times6}$$

$$nCr = \frac{n(n-1)(n-2) \times \ldots\ldots\ldots(n-r+1)}{1 \times 2 \times 3 \times 4 \times \ldots\ldots\ldots r.}$$

The above principle can be used in our future calculations.

Example 1. A ball is drawn from a bag containing 5 white balls and 7 red balls. What is the probability that the ball drawn is a white ball ?

Since there are 5 white balls, the white balls can be drawn in 5 ways. Therefore, the event can occur in 5 ways.

As the total number of balls is 12, a ball can be drawn in $12C_1$ ways i.e : 12 ways.

∴ The probability of getting a white ball

$$= \frac{\text{Number of favourable cases}}{\text{Number of possible cases}}$$

$$= \frac{5}{12}$$

2. Find the probability of having a king from a packet of playing cards.

Total number of cards = 52.

Total number of kings = 4.

Total number of possible cases $52C_1$ = 52.

Number of favourable cases = $4C_1$ = 4.

$\therefore$ Probability $= \frac{4}{52} = \frac{1}{13}$

3. Two cards are drawn from a pack of 52 cards. Find the chance of getting two queens.

Total number of possible cases of drawing 2 cards from 52 cards = $52C_2$

$$= \frac{52 \times 51}{1 \times 2} = 26 \times 51$$

Total number of queens = 4

Possible ways of getting 2 queens $= 4c_2 = \frac{4 \times 3}{1 \times 2} = 6.$

Possibilities of getting 2 queens $= \frac{6}{26 \times 51} = \frac{2}{26 \times 17}$

$$= \frac{1}{221}$$

Additional Theorem of Probability

If two events E_1 and E_2 are mutually exclusive, then the probabilitv for the occurrence of either E_1 or E_2 can be written in the symbolic form $P(E_1+E_2) = P(E_1) + P(E_2)$

Let us assume that out of a total of 'n' causes, m_1 causes favour the occurrence of the event E_1 and m_2 causes favour the occurrence of the event E_2. The probability for the occurrence of the event either E_1 or E_2 will be equal to the sum of the probabilities for the occurrence of the events E_1 and E_2.

$$\text{Probability of } E_1 = P(E_1) = \frac{m_1}{n}$$

$$\text{Probability of } E_2 = P(E_2) = \frac{m_2}{n}$$

$$\text{Probability of } E_1 \text{ or } E_2 = P(E_1+E_2) = P(E_1)+P(E_2)$$

$$= \frac{m_1}{n} + \frac{m_2}{n} = \frac{m_1 + m_2}{n}$$

General formula

Example : $P(E_1+E_3+\ldots\ldots) = P(E_1) + P(E_2) + P(E_3)+\ldots\ldots$

A bag contains 3 red balls 4 green balls and 5 yellow balls.

Total No. of balls = 3 + 4 + 5 = 12.

The probability of taking 1 red ball $= \frac{3}{12} = P(E1)$

The probability of taking 1 green ball $= \frac{4}{12} = P(E2)$

The probability of taking 1 yellow ball $= \frac{5}{12} = P(E3)$

1. Probability of taking either a red or a green ball

$$P(E1 + E2) = \frac{7}{12} = \frac{3}{12} + \frac{4}{12} = P(E1) + P(E2)$$

2. Probability of taking either a red or yellow ball $= \frac{8}{12}$

$$P(E1 + E3) = \frac{3}{12} + \frac{5}{12} = P(E1) + P(E3)$$

3. Probability of taking either a green or yellow ball $= \frac{9}{12}$

$$P(E2 + E3) = \frac{4}{12} + \frac{5}{12} = P(E2) + P(E3).$$

4. Probability of taking either a red, or green or yellow ball $= \frac{12}{12}$

$$P(E1 + E2 + E3) = \frac{3}{12} + \frac{4}{12} + \frac{5}{12}$$

$$= P(E1) + P(E2) + P(E3).$$

It should be remembered in this case that the occurrences of red ball, green ball and yellow ball are mutually exclusive; since if the ball drawn is a red ball; the event of drawing a green ball is not affected.

Let us now consider the cases where the occurrences are not mutually exclusive. Take a pack of cards containing 10 cards out of which 5 cards are red and 5 are black. Again out of these, 2 cards in each colour are containing pictures and the remaining 3 in each colour are containing only numbers. We can represent this in the following table.

Colour	Picture	Number	Total
Red	2	3	5
Black	2	3	5
	4	6	10

The probability of getting a red card is $\frac{5}{10}$ P(E1)

The probability of getting a picture card is $\frac{4}{10}$ P (E2)

The simultaneous occurrence of getting red and picture card $= \frac{2}{10} = P\ (E1\ E2)$

Therefore the probability of getting either red or picture card $= \frac{5}{10} + \frac{4}{10} - \frac{2}{10}$

$$= P(E1) + P(E2) - P(E1\text{-}E2)$$

The 2 red picture cards are included in both the cases ($\frac{5}{10}$ & $\frac{4}{10}$) and hence it has to be subtracted once.

$$\frac{7}{10} = \frac{5 + 4 - 2}{10}$$

Therefore $P(E1 + E2) = P(E1) + P(E2) - P(E1E2)$.

1. If two events are such that the sum of their probabilities is equal to 1, the events are said to be complementary.

2. If the two events are mutually exclusive, then $P(E1 + E2) = P(E1) + P(E2)$.

3. If the events are mutually not exclusive, then

$$P(E1 + E2) = P(E1) + P(E2) - P(E1\,E2).$$

Multiplication theorem

Two events are said to be independent when the occurrence of one event say E1 does not affect in any way the occurrence of the other event E2.

If two independent events E1 and E2 are given, then the probability of the simultaneous occurrence of E1 and E2 will be equal to the product of their probabilities.

Events	Probability
E1	P (E1)
E2	P (E2)
E3	P (E3)

$$P\,(E1\ E2\ E3) = P\,(E1) \times P\,(E2) \times P\,(E3)$$

Example : Find out the probability of getting 3 heads in tossing 3 coins at a time or tossing a single coin 3 times. Here the events are independent since the result obtained in one throw does not affect the result in the other throw. The probability of getting a head in tossing a coin $= \frac{1}{2}$.

The probability of getting 2 heads at a time will be $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

The probability of getting 3 heads at a time will be $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$.

Conditional Probability

If two events E1 and E2 are given, then the probability of the simultaneous occurrence of E1 and E2 is equal to the product of the probability of E1 and the conditional probability of E2 given the probability of E1.

Three bags, A, B and C contain white and red balls in the following manner :

Bags	White W	Red R	Total
A	1	1	2
B	2	...	2
C	...	2	2
Total	3	3	6

A bag is chosen at random and a ball is taken out. What is the probability that the ball extracted is a white ball?

In order to get a white ball we should take either the Bag A or Bag B.

Let us consider Bag A

The probability of getting the Bag A $= \frac{1}{3} = P(A)$

The probability of getting a white ball out of Bag A

$$P\left[\frac{W}{A}\right] = \frac{1}{2}$$

Therefore, the probability of getting a white ball from Bag A $= \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$ i.e. P(AW)

Similarly the probability of getting a white ball from Bag B $= \frac{1}{3} \times \frac{2}{2} = \frac{1}{3}$ i.e. P(BW).

Therefore the probability of getting a white ball

$$= \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

Addition and Multiplication of Probabilities

Let us take 2 dice and find out the probability of getting a total of 6 in throwing the two dice simultaneously. Let the two dice be indicated by the letters A and B. The total 6 can be obtained in the following manner.

Dice	*Number in the faces*				
A	1	2	3	4	5
B	5	4	3	2	1
	6	6	6	6	6

The chance of getting the number 1 in the first die is $\frac{1}{6}$. Similarly the chance of getting 5 in the second die is $\frac{1}{6}$. Therefore, the chance of getting simultaneously 1 in the first die and getting 5 in the second die is equal to

$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Similarly in the case of other combination also the probability is $\frac{1}{36}$ as given below :

A	B	
1	5	$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
2	4	$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
3	3	$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
4	2	$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
5	1	$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

The probability of getting a total of 6 in throwing the two dice will be equal to the sum total of the individual probability.

$$\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{5}{36}$$

(Eg.) Let us find that the probability of getting a sum total of the faces is equal to at least 6 ie: 6 or above in throwing 2 dice. The sum totals can be either 6 or 7 or 8 or 9 or 10 or 11 or 12.

1. Probability of getting 6 = A = 1 2 3 4 5

$$B = \frac{5}{6}\ \frac{4}{6}\ \frac{3}{6}\ \frac{2}{6}\ \frac{1}{6}$$

$$= 5 \times \frac{1}{36} = \frac{5}{36}$$

2. Probability of getting 7 = A 1 2 3 4 5 6
B 6 5 4 3 2 1
7 7 7 7 7 7

$$= 6 \times \frac{1}{36} = \frac{6}{36}$$

3. Probability of getting 8 = A 2 3 4 5 6
B 6 5 4 3 2

$$= 5 \times \frac{1}{36} = \frac{5}{36}$$

4. Probability of getting 9 = A 3 4 5 6
B 6 5 4 3
9 9 9 9

$$= 4 \times \frac{1}{36} = \frac{4}{36}$$

5. Probability of getting 10 = A 4 5 6

B 6 5 4

$$= 3 \times \frac{1}{36} = \frac{3}{36}$$

6. Probability of getting 11 = A 5 6

B 6 5
11 11

$$= 2 \times \frac{1}{36} = \frac{2}{36}$$

7. Probability of getting 12 = A 6

B 6
12

$$= 1 \times \frac{1}{36} = \frac{1}{36}$$

8. Probability of getting a sum atleast equal to 6 or above

$$\frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36}$$

$$= \frac{26}{36} = \frac{13}{18}$$

BINOMIAL EXPANSION AND BINOMIAL DISTRIBUTION

The statistical data collected are presented in the form of a frequency distribution. These distributions are based on actual data. But there are certain distributions which are not based on actual data or experiments, but they are derived mathematically or theoretically on the basis of certain assumptions. Hence these distributions are called, theoretical distributions. They may also be called as Expected Frequency Distributions since the frequencies for the different values are not actuals but are expected frequencies according to mathematical base of the theory. There are three types of such distributions such as Binomial Distribution, Normal Distribution and Poisson Distribution. We shall discuss the Binomial distribution first.

Binomial Distribution

Let us consider the tossing of 2 coins simultaneously and find out the possible occurrences.

Coin A : H H T T

Coin B : H T H T

Where H denotes Head and T denotes Tail. This can be written as follows :

HH, HT, TH, TT.

The chance of getting a head or tail in a single coin is equal to $\frac{1}{2}$. Therefore the probabilities for the above occurrences are as follows :

Occurrences	*Probabilities*
HH	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
HT	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
TH	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
TT	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

In the above case, the chance of getting 2 heads is equal to $\frac{1}{4}$, the chance of getting atleast one head or atleast one tail is equal to $2 \times \frac{1}{4} = \frac{1}{2}$ and the chance of getting 2 tails is equal to $\frac{1}{4}$.

Let us consider the appearance of 'head' as a success and denote its probability as P. Then the appearance of tail will be considered as a failure and its probability is denoted by q so that $p + q = 1$ or $q = 1 - p$.

The occurrence of heads and tails as given above can be written in terms of the probabilities as follows :

HH	HT	TH	TT
pp	pq	qp	qq
p^2	pq	qp	q^2
p^2	(pq + pq)		q^2
p^2	2 pq		q^2

$(p^2 + 2pq + q^2)$ is the product obtained from the expansion of the term $(p + q)^2$.

From this we can infer that if 2 events are independent then their combined probability can be given by the expansion of the term $(p + q)^2$.

If $p = \frac{1}{2}$ and $q = \frac{1}{2}$ then the probabilities of the various events can be given as follows :

2 heads $p^2 = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

1 head and 1 tail $= 2pq = 2 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$

2 tail $= q^2 = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

The results can be given in the following form. The expression can be $(\frac{1}{2} + \frac{1}{2})^2 = (q + p)^2$

Number of success	*Probability*
0	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
1	$2 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$
2	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Let us now consider 3 coins and all possible outcomes :

H	H	H	$= p \times p \times p = p^3$	p^3
H	H	T	$= p \times p \times q = p^2q$	
H	T	H	$= p \times q \times p = p^2q$	$3p^2q$
T	H	H	$= q \times p \times p = p^2q$	

T H T	$= q \times p \times q = pq^2$	}	$3pq^2$
T T H	$= q \times q \times p = pq^2$		
H T T	$= p \times q \times q = pq^2$		
T T T	$= q \times q \times q = q^3$		q^3

The result $p^3 + 3p^2 q + 3pq^2 + q^3$ is nothing but the expansion of the form $(p + q)^3$.

If the values of p and q are given we can get the probability of the various events.

It may be seen from the above that in general we can use the form $(p + q)^n$ or more general by $(q + p)^n$ where n stands for the number of coins or the number of tosses with a single coin or the number of trials with a single coin.

Probability of compound events : The above result can be explained as follows. If the probability for an occurrence of an event is p and the probability of its failure is q, $(1—p)$, then the probability for occurring in 'r' times out of a total 'n' coins can be written in the form

$$_nC_r \, p^r \, q^{n-r} \text{ where } q = 1 - p$$

$$\text{and } _nC_r = \frac{\underline{|n}}{\underline{|r} \times \underline{|n-r}} = \frac{1 \times 2 \times 3 \times 4 \times \ldots\ldots\ldots \times n}{(1 \times 2 \times 3 \times \ldots r) \times (1 \times 2 \times 3 \times \ldots n\text{-}r)}$$

If the event occurs 'r' times it means it fails in $(n—r)$ times. According to the law of Multiplication, the probability for occurring r times and failing $(n - r)$ times will be $p^r \times q^{n-r}$.

It may be any 'r' occasions starting from any occasion out of the 'n' occasions. This is equal to taking a combination of 'r' occasions out of a total of 'n' occasions and the result can be symbolically written as $_nC_r$ ways, which means the number of ways in which a combination of 'r' items taken out of 'n' items. As there are $_nC_r$ ways, according to the law of addition the probability is $_nC_r \, p^r \, q^{n-r}$ which is the general term of Binomial expansion $(p+q)^n$. But generally, it is written in the form $(q+p)^n$. In order to find the probability for the number of occurrences starting from 0, 1, 2, ... the expression is written as $(q+p)^n$.

Expansion

The Binomial Expression is $(q+p)^n$ and its expansion is as follows :

$$(q+p)^n = q^n + {}_nC_1 q^{n-1} p + {}_nC_2 q^{n-2} p^2 + {}_nC_3 q^{n-3} p^3 + \dots \dots\dots + {}_nC_r q^{n-r} p^r + \dots\dots + {}_nC_n p^n$$

As before, the number of success and their respective probabilities can be given in a tabular form as follows :

Number of success	Probability
0	${}_nC_o q^n p^o = {}_nC_o q^n = q^n$
1	${}_nC_1 q^{n-1} p$
2	${}_nC_2 q^{n-2} p^2$
3	${}_nC_3 q^{n-3} p^3$
...	
...	
r	${}_nC_r q^{n-r} p^r$
...	
...	
(n-1)	${}_nC_{(n-1)} q p^{n-1}$
n	${}_nC_n q^0 p^n$ or ${}_nC_n p^n = p^n$

The above table is called Binomial frequency distribution and in short Binomial distribution. It can be defined as follows :

If an experiment consisting of 'n' trials is conducted with a probability of the success of a particular event in each trial is equal to 'p' and the probability of its failure in each trial is 'q' such that $p+q = 1$, then the probability of getting 0, 1, 2, 3, ..., n success can be given by the successive terms in the expansion of $(q+p)^n$.

If the above experiment, each consisting of 'n' trials is repeated N times, then the number of experiments in which we get 0, 1, 2, ..., n success are given by the successive terms in the expansion of $N(q+p)^n$.

The Binomial distribution is based on the following two assumptions :

1. The events should be discrete (such as 0, 1, 2, 3, 4, ...) and not continuous such as 1·5, 2·3, 3·1 etc.

2. The probability of its success in each trial 'p' shall be the same. In other words, the probability of its failure $q = (1 - p)$ in each trial should be the same. It means, the values of p and q should be constant throughout.

Example

A coin is tossed 5 times. Find the probability of getting exactly 2 heads and 3 tails.

Since the coin is tossed 5 times, $n = 5$.

Let the occurrence of head be considered success 'p' and tail as its failure 'q'.

The probability of getting a head in a single trial $= \frac{1}{2} = p$.

$\therefore \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$.

We have to find the probability of getting 2 heads and 3 tails. Probability of getting 'r' success in 'n' trials $= {}_nC_r \; q^{n-r} \; p^r$. But in the problem $n = 5$, $r = 2$, $p = \frac{1}{2}$, $q = \frac{1}{2}$.

$\therefore$ Probability of getting 2 success

$$= {}_5C_2 \, p^2 \, q^3$$
$$= {}_5C_2 \left(\tfrac{1}{2}\right)^2 \left(\tfrac{1}{2}\right)^3$$
$$= {}_5C_2 \left(\tfrac{1}{2}\right)^5$$
$$= \frac{5 \times 4}{1 \times 2} \times \left(\frac{1}{2}\right)^5$$
$$= 10 \times \left(\tfrac{1}{2}\right)^5$$
$$= 10 \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2}$$
$$= \frac{5}{16}$$

Example

Number of coins $= n = 4$

Number of heads required $= 2$

$p = \frac{1}{2}, \ q = \frac{1}{2}$

Probability of getting 2 heads :

$${}_4C_2 \left(\tfrac{1}{2}\right)^2 \left(\tfrac{1}{2}\right)^2 = \frac{4 \times 3}{1 \times 2} \times \left(\tfrac{1}{2}\right)^4$$

$$= 2 \times 3 \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \tfrac{3}{8}$$

Mean and Standard Deviation of Binomial Distribution

Mean = np.

Standard Deviation = $\sqrt{npq}$

The derivation of these results is beyond the scope of this book.

Example

The probability of a defective bulb is *0·01*. Find the Mean and Standard Deviation of defective bulbs in a total of *100,000* bulbs. n = *100,000*, p = *0·01*, q = (1—p)

$$q = (1 - 0{\cdot}01) = 0{\cdot}99$$

Mean = np = *100,000* × *0·01* = **1000** bulbs.

Standard Deviation $= \sqrt{npq}$

$= \sqrt{100{,}000 \times 0{\cdot}01 \times 0{\cdot}99}$

$= \sqrt{990}$

$= \sqrt{9{\cdot}9 \times 100}$

$= 3{\cdot}146 \times 10 = \mathbf{31{\cdot}5}$

Characteristics of Binomial Distribution

1. The general form of Binomial distribution depends upon the values of n, p and q.

2. If the values of p and q are equal, the Binomial distribution will be symmetrical.

3. If p and q are not equal it will be a skewed distribution.

4. Even if p and q are not equal, the distribution will tend to be symmetrical if 'n' is very large.

5. The Mean and the Standard Deviation will be equal to 'np' and $\sqrt{npq}$ respectively.

Normal Distribution

The Binomial distribution is a discrete distribution. When $p = q = \frac{1}{2}$ and 'n' becomes infinitely large, the Binomial distri-

bution tends to be a symmetrical and continuous distribution called Normal distribution.

Normal distribution can be expressed graphically. The graph of Normal distribution is known as Normal curve or Normal probability curve. The equation to a Normal curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x - \bar{x})^2}{2\sigma^2}}$$

where $\bar{x}$ is the Mean; σ is the Standard Deviation.

$\pi = 3{\cdot}1428 \quad e = 2{\cdot}71828$

y = the ordinate or height of the curve at a point at a distance x from the origin.

Generally the equation will be written in the form

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

where $t = \frac{(x - \bar{x})}{\sigma}$

If N = 1, this equation will be of the form

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

This is the standard form of the curve and it will have a bell shape as given below :

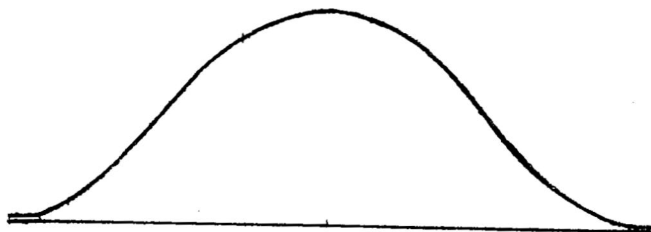


Fig. 1 - 1

When N = 1, the total area under the curve is unity.

Let us take two points at a distance of t_1 and t_2 from the origin. If the ordinates are drawn at these points so as to touch the curve, the area under the curve and between the two ordinates will represent the probability that the value of 't' lies between t_1 & t_2.

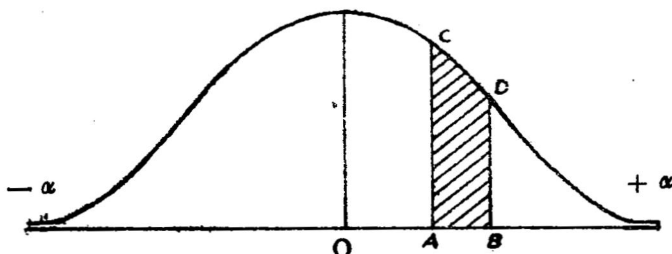


Fig. 1 - 2

Since the curve is symmetrical starting from $-\alpha$ to $+\alpha$ the centre 'O' is taken as the origin. Let $OA = t_1$, $OB = t_2$. AC and BD are the two ordinates at A and B at a distance of t_1 and t_2 from the origin.

The area under the curve and the ordinates =
The area covered by the portion ABCD.

The following table gives the area under the normal curve and the ordinate drawn at a distance of 't' from the origin. This is obtained by using the following formula

$$\int_{-\alpha}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Negative Value of 't'	Area	Positive Value of 't'	Area
		0·0	0·5000
—0·1	0·4602	0·1	0·5398
—0·2	0·4207	0·2	0·5793
—0·3	0·3821	0·3	0·6179
—0·4	0·3446	0·4	0·6554
—0·5	0·3085	0·5	0·6915
—0·6	0·2743	0·6	0·7257
—0·7	0·2420	0·7	0·7580
—0·8	0·2119	0·8	0·7881
—0·9	0·1841	0·9	0·8159
—1·0	0·1587	1·0	0·8413
—1·1	0·1357	1·1	0·8643
—1·2	0·1151	1·2	0·8849
—1·3	0·0968	1·3	0·9032
—1·4	0·0808	1·4	0·9192
—1·5	0·0668	1·5	0·9332
—1·6	0·0548	1·6	0·9452
—1·7	0·0446	1·7	0·9554
—1·8	0·0359	1·8	0·9641
—1·9	0·0287	1·9	0·9713
—2·0	0·0228	2·0	0·9772
—2·1	0·0179	2·1	0·9821
—2·2	0·0139	2·2	0·9861
—2·3	0·0107	2·3	0·9893
—2·4	0·0082	2·4	0·9918
—2·5	0·0062	2·5	0·9938
—2·6	0·0047	2·6	0·9953
—2·7	0·0035	2·7	0·9965
—2·8	0·0026	2·8	0·9974
—2·9	0·0019	2·9	0·9981
—3·0	0·0013	3·0	0·9987

With the help of the above table we can get the probability that a value lies between two given values when the Mean and Standard Deviation of the distribution are given.

Example

(1) A normal distribution has the following details :

Mean = 20; Standard Deviation = 4.

Find the probability that a value x lies between 20 & 24.

We know $t = \frac{x - \bar{x}}{\sigma}$

When x = 20, $t = \frac{20-20}{4} = \frac{0}{4} = 0 = t_1$

When x = 24, $t = \frac{24 - 20}{4} = 1 = t_2$

When x lies between 20 and 24 it means 't' lies between 0 and 1. Therefore, it is required to find the probability that 't' lies between 0-1. This probability is given by the area under the curve x axis and the 2 ordinates drawn at a distance 0 and 1 from the origin.

It may be seen from the above table that the area is equal to 0·8413 corresponding to the value of t = 1. Therefore, the probability is equal to 0·8413.

The area corresponding to the value of t = 0 is 0·5000. Therefore the probability is 0·5000. Hence the probability for the value to be between 20 and 24 = 0·8413 — 0·5000

= 0·3413

Let us consider the following and calculate the probability for the values lying

(i) below 12

(ii) between 12 & 16

(iii) between 16 & 28

In the above example,

Mean $\bar{x} = 20$

(i) $t = \frac{x - \bar{x}}{\sigma} = \frac{12 - 20}{4} = -2$

The probability for the positive value of $t = 2$

$= 0{\cdot}9772.$

$\therefore$ The probability for the negative value of $t = -2$

$= 1 - 0{\cdot}9772 \quad = 0{\cdot}0228.$

(ii) $t_1 = \frac{x_1 - \bar{x}}{\sigma} = \frac{12 - 20}{4} = -2$

$t_2 = \frac{x_2 - \bar{x}}{\sigma} = \frac{16 - 20}{4} = -1$

The probability for the negative value of

$t_1 = -2 = 1 - 0{\cdot}9772 = 0{\cdot}0228$

The probability for the negative value of

$t_2 = -1 = 1 - 0{\cdot}8413 = 0{\cdot}1587$

The probability for the value lying between

t_1 & $t_2 = 0{\cdot}1587 - 0{\cdot}0228 = 0{\cdot}1359$

(iii) $t_1 = \frac{x_1 - \bar{x}}{\sigma} = \frac{16 - 20}{4} = -1$

$t_2 = \frac{x - \bar{x}}{\sigma} = \frac{28 - 20}{4} = 2$

The probability for the negative value of $t_1 = -1$

$= 1 - 0{\cdot}8413$

$= 0{\cdot}1587$

The probability for the positive value of $t_2 = 2$

$= 0{\cdot}9772$

The probability of t for lying between t_1 & t_2

$= 09772 - 0{\cdot}1587$

$= 0{\cdot}8185$

Example

In the above example find the probability for the value of x to lie between 24 and 28.

When $x = 24$, $t = \frac{x - \bar{x}}{\sigma}$

$$= \frac{24 - 20}{4} = 1 = t_1$$

When $x = 28$, $t = \frac{x - \bar{x}}{\sigma} = \frac{28 - 20}{4} = 2 = t_2$

Probability that x lies between 24 and 28 is the same as the probability for 't' lying between 1 and 2.

The probability for 't' lying below 2 = 0·9772

The probability for 't' lying below 1 = 0·8413

∴ The probability for 't' lying between 1 - 2 = 0·9772 — 0·8413 = 0·1359

∴ The probability of x lying between 24 and 28 = **0·1359**

When we know the total frequency (N) and the probability, we can also determine the frequency by multiplying the total frequency by the probability.

The total frequency of a normal distribution is equal to 1000. Its mean is 35 and the standard deviation is 7. Find the frequency of the value lying between 42-49.

$$N = 1000, \ \bar{x} = 35, \qquad \sigma = 7.$$

Value of 't' when $x = 42$ is $\frac{42 - 35}{7} = 1$

Value of 't' when $x = 49$ is $\frac{49 - 35}{7} = 2$

The probability for 't' lying below 2 or x lying below $49 = 0{\cdot}9772$

The probability for 't' lying below 1 or x lying below $42 = 0{\cdot}8413$

$\therefore$ The probability of x lying between 42 and 49

$$= 0{\cdot}9772 - 0{\cdot}8413$$
$$= 0{\cdot}1359$$

$\therefore$ The frequency $= 1000 \times 0{\cdot}1359$
$= 135{\cdot}9 \doteq 136$

Properties of Normal Curve

1. The normal curve is a unimodal, symmetrical and perfectly bell shaped. The ends of the curve get closer and closer to the X-axis as we move from the Mean but they never touch the X-axis.

2. The Mean and Median coincide with the Mode.

3. The total area under the normal curve is equal to the total frequency. The ordinate drawn through a point at a distance equal to the mean of the distribution from the origin divides the total area under the curve into two equal parts.

4. Co-efficient of skewness $= 0$.

5. Measure of Kurtosis $\beta_2 = 3$.

6. The two quartiles are equidistant from the Median.

7. About 68% of the total items lies between $\bar{x} - 1\sigma$ and $\bar{x} + 1\sigma$.

8. About 95% of the total items lies between $x - 2\sigma$ and $\bar{x} + 2\sigma$.

9. About 99% of the total items lies between $x - 3\sigma$ and $\bar{x} + 3\sigma$.

The following diagram will explain these trends.

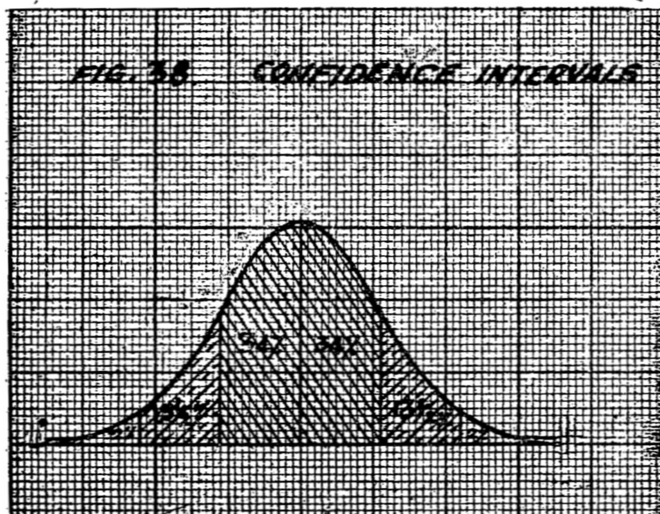


Fig. 1 - 3

POISSON DISTRIBUTION

We have studied earlier that the Binomial distribution would be a Normal distribution in the limiting case even if p and q are unequal provided 'n' was increased sufficiently large. We shall now see that the limit of the same distribution when 'p' is small and 'n' is large, so that 'np' is finite.

This is known as Poisson series or Poisson exponential limit.

However, the practical value of this series is very limited. Hence the examples of this distribution are generally called "rare events" as 'p' the probability of occurrence is small. In practice, variables like the number of motor accidents in a city per day, which can take as big a value as the whole population of the city, but ordinarily it is only a small number, follows frequency distribution of this pattern.

It can be further explained. If the probability of the occurrence of an event 't' in a single trial be a small quantity 'p' and 'n' trials are performed, where 'n' is sufficiently large to make 'np' a constant equal to 'm', the probability of the event occurring exactly 'x' times is given by

$$P(x) = \frac{e^{-m} m^x}{1 \times 2 \times 3 \times \ldots\ldots x}$$

We know that the probability for 'x' success in 'n' trials in a Binomial distribution is,

$P(x) = {}_nC_x \, p^x q^{n-x}$ when $n \rightarrow \alpha$, with 'np' constant.

If we put $p = \frac{m}{n}$

we get, $\frac{e^{-m} m^x}{\underline{|x}}$

where $\underline{|x} = 1 \times 2 \times 3 \times \ldots\ldots x$

Sum of all the probabilities

The value of x and its probability in a Poisson distribution are as follows :

Value (x)	Probability
0	$\frac{e^{-m} m^0}{\underline{|0}}$
1	$\frac{e^{-m} m^1}{\underline{|1}}$
2	$\frac{e^{-m} m^2}{\underline{|2}}$
...	...
...	...
x	$\frac{e^{-m} m^x}{\underline{|x}}$

The sum of all the probabilities is equal to 1.

The sum of all the probabilities

$$= e^{-m}\left(1 + \frac{m}{\underline{|1}} + \frac{m^2}{\underline{|2}} + \frac{m^3}{\underline{|3}} + \ldots\ldots\right)$$

i. e. $e^{-m} \times e^{m} = 1$

$$e^{m} = 1 + \frac{m}{\underline{|1}} + \frac{m^2}{\underline{|2}} + \frac{m^3}{\underline{|3}} + \ldots\ldots$$

Constants

Poisson distribution contains only one parametre namely 'm'. The estimate of 'm' is furnished by the simple Arithmetic Mean.

Mean and the Variance

In a Poisson distribution, the variance is equal to the Arithmetic Mean (m) and this fact is used to test whether a given distribution follows the Poisson Law.

Example

The classical example of a Poisson distribution gives the frequency of the number of deaths due to kick of a horse in 10 corps per army per annum over twenty years.

x	f
0	109
1	65
2	22
3	3
4	1
Over 4	0
	200

Let us calculate now $\bar{x}$ and σ^2

x	f_i	$x_i f_i$	$x_i^2 f_i$
0	109	0	0
1	65	65	65
2	22	44	88
3	3	9	27
4	1	4	16
Over 4	0	0	0
	200	122	196

$$\bar{x} = m = \frac{122}{200} = 0{\cdot}61$$

$$\text{Variance} = \frac{196}{200} - m^2$$

$$= \frac{196}{200} - 0{\cdot}61 \times 0{\cdot}61$$

$$= 0{\cdot}98 - 0{\cdot}3721$$

$$= 0{\cdot}6079$$

or **0·61**

Exercise

1. Define probability.
2. Explain the additional theorem of probability with an example.
3. Explain the Multiplication theorem of probability with an illustration.
4. Find the probability that the sum of the numbers will be 10 in a throw of 2 dice.
5. An urn contains 5 red and 10 green balls. What is the probability that 8 balls drawn in succession will give 3 red and 5 green balls ?

6. A & B each chooses a digit from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Find the probability

   (a) that the sum of the two digits is (i) 15 (ii) 10 (iii) 12

   (b) Product of the digits is (i) 24 (ii) 54 (iii) 36

7. Find the Binomial distribution whose mean is 15 and Standard Deviation is $\sqrt{6}$.

8. A man tosses a coin 8 times. What is the probability of getting

   (i) all heads

   (ii) 6 heads and 2 tails

   (iii) 5 heads or less.

9. Give the mean and variance of a Poisson distribution.

10. State the properties of normal distribution.

CHAPTER II

SAMPLE SURVEYS

We have studied earlier how the statistical details are collected. We have also seen the different methods for collection of statistical details and also the merits and demerits of collecting data through Correspondence method, Registration method, Census method etc.

Origin of Sampling

In practical problems the statistician is confronted with the necessities of discussing a universe of which he cannot examine every member. Perhaps in the process of examining the characteristics of the universe, the universe may be destroyed. In such cases the best that an investigator can do is to examine a limited number of individuals or units or items and hope that they will represent the universe as much as he wants to know about the universe from which the individuals came. Such limited number of units from the universe may be called samples from the universe. This is the origin of the theory of sampling.

Complete Enumeration

As explained earlier the important function of statistician is collection of statistics. One way of collecting data is by the process of complete enumeration. This consists of knowing all the units about which information is required and collecting the information for all such units. The population census and the livestock census conducted in our country are suitable examples for complete enumeration.

Theory of Sampling

A sample from a universe is a selected number of units or individuals each of which is a member of the sample. The uni-

verse can be divided into 2, namely finite and infinite universe. A finite universe is a universe which contains finite number of individuals while the infinite universe is one with an infinite number of units or individuals. A hypothetical universe can be defined as the aggregate or total of all conceivable ways in which specified event or incident can happen. The infinite number of throws which can be made by a coin or die can be classified under this.

Advantages of Sample Survey

An alternative to complete enumeration is a sample survey. Here data are collected only from a few of the units that would be included in a complete enumeration. Collection and compilation of a small volume of data need less number of men and also is less expensive. Because of this, training and organising a machinery with a small staff will not involve much difficulties. In view of these advantages, such surveys can be repeated at frequent intervals and built a chain of information. This will also ensure in building a well trained human machinery for purposes of future surveys. There are occassions where complete enumeration is also not possible and in all such cases only sample surveys have to be adopted. Even the accuracy of the information collected in complete enumeration can be ensured only by means of sample checks. By means of sample surveys, advance estimates can be made.

In all sample surveys the results obtained are only estimates and not absolute values. The error in the estimates made with the help of sample surveys can also be estimated by adopting suitable sampling techniques. In certain cases the error in the estimate can also be approximately estimated in advance. Generally, the administrator may not be interested in knowing the exact actuals. Instead, he would be satisfied if he is supplied with the result with a reasonable margin of error for the purpose of taking decisions on policies. Sample surveys are best tools in such circumstances. However, sample surveys may not be useful where information regarding each and every member of the universe are required.

As said earlier, an aggregate of units is termed as population in statistical terminology and each unit in the population

is called a sample unit. Such s mple units may be natural units or artificial units. The sample unit may not be always of uniform size. However, sample unit must be clearly and unambiguously defined for a particular survey.

The very object of sample survey is to know an estimate of the population value with a reasonable margin of error with less cost and at the same time within a limited period. This object can be best achieved only if specific survey is adopted in specific cases. Therefore, different types of surveys have been evolved by statisticians and each method has its own merits and demerits.

Sampling Frame

A sampling frame is a description of all the sample units which constitute the population. It is the basis for drawing sample. The sampling frame may be a map or a list or any other source from which a few of the sample units can be drawn according to the design of the survey.

The fundamental object of sampling is to obtain maximum information about the 'parent universe' with the minimum effort. The process of forming a sample consists of choosing a predetermined number of individuals from the parent universe. This can be done in three ways, namely (1) by selecting the individuals at random (2) by selecting the individuals according to some purposive principles and (3) by a combination of the above two methods.

Random Sampling

A definition of random sampling may be given by saying that the selection of an individual from a universe is random when each and every member of the universe has the same or equal chances of being selected or chosen. A sample of 'n' individuals is random when it is chosen in such a way that when the choice is made, all possible samples of size 'n' units have an equal chance of being selected.

The problem of obtaining a random sample is more difficult than it appears at first sight. Any purely haphazard

method of selection will not give a random sample. The method of selection must follow some code of procedure which will leave nothing to the observer's idiosyncrasy. Whenever there is scope for personal joy or judgement on the part of the observer, bias is almost certain to creep in. This bias cannot be removed by any effort because human being has always a tendency to be away from true randomnes in his choice.

The criterion that every individual must have an equal chance of being selected may be modified. If the method of selection is independent of the properties of the sample universe, there will be no reason why one individual should be chosen rather than the other. Hence all values of the properties which occur in the universe will have, an equal chance of being chosen. If, therefore, a mode of procedure which bears no relationship to the properties of the parent universe can be devised, it may be expected that the sample chosen will be a random one. Thus, if the members of a given universe are serially numbered and a sample is chosen by selecting the individuals corresponding to numbers at constant intervals beginning with an arbitrary start, it may be expected that the sample chosen may be a random sample. This method will fail if certain characteristics of the universe repeat at the same intervals.

Miniature Universe

One of the most reliable methods of choosing a random sample is by choosing it from a miniature universe, whose members exhibit a one to one correspondence with the members of the original universe. This miniature universe may consist of pieces of paper or small similar balls of same material, same size and shape on which the numbers corresponding to members of the original universe are written. The pieces of papers or balls are placed in similar containers, usually metal cylinders and are thrown into a large rotating drum in which they are thoroughly mixed or randomised. Afterwards the required number of individuals are taken from this drum as in the case of lotteries.

If the original universe is large, there will be lot of practical difficulties in constructing a miniature universe and shuffling. In such cases, use of random numbers are adopted to select the samples.

Random Numbers

Random numbers have been constructed by many statisticians including Tippet, Fisher and Yates. All are published random numbers. These random numbers ensure that the digits 0—9 occur equally, frequently in horizontal vertical or diagonal directions. Also, combination of digits 00 to 99 occur equally frequently and so on. The sequence in which the digits occur do not follow any law. A set of random numbers containing one digit, two digits and three digits are given in the Appendix for our use.

The numbers are being chosen by really random methods. But there is no proof except by actual experience to say that these numbers are random. Thus to select a sample of 'n' individuals from a population of 'N' individuals, the individuals in the population are numbered serially from 1 to N. The procedure then is to take any page of the random numbers and choose the first 'n' numbers occurring on that page after rejecting numbers greater than 'N'. The individuals corresponding to these 'n' random numbers chosen will constitute a random sample of 'n' units.

Continuous Universe

However, a different technique will have to be adopted in drawing a sample from continuous population like a barrel of flour or a bag of rice. One method will be to divide the population (bag of rice) into a large number of small packets of equal size and then take a random sample of packet after giving serial numbers to the packets. Sometimes the flour may be thoroughly mixed and divided further into two equal halves and one half of it is chosen at random and the process is repeated while a suitable sample of required size is obtained.

Hypothetical Population

A random sample from a hypothetical population like the universe of throws of a die or coin is obtained by throwing a die or coin into the required number of times and taking the result as a sample. Care must also be taken so that sampling conditions remain constant throughout the experiment.

A random sample gives a quite satisfactory estimate of the parent universe when the population is more or less homogeneous. But, when the parent universe is hetrogeneous and the number of members in the sample is small, the sample may often give incorrect estimates of the universe. To remedy this kind of trouble, typical representative members of the population are chosen and the sample obtained by this method is known as purposive sample. It may be noted that as the size of the sample or number of members in the sample increases the random sample will give a better approximation of the universe than the purposive sample. The object of the sampling in many cases is to get the information about the whole universe. Hence the objection to the purposive sampling is that it may give a better result about the typical members of the universe and probably it would give a poor idea of the degree of variance of the characteristics of the members of the universe.

In many types of statistical investigations, a combination of two methods of sampling is used to get a satisfactory result. This is particulary proved if the structure of the parent universe is practically known.

Random Sample

A method of selection is said to be random if every unit in the aggregate of units or population or universe has an equal chance of being selected. In random sample, a definite number of units are chosen according to the laws of chance.

Let us suppose that we have to select a random sample of 'n' distinct units from a population of 'N' plots. While

drawing the first plot, the random process should be such that every plot in the whole N plots has an equal chance of being chosen. After the selection of the first plot, the second plot can be chosen out of the remaining (N-1) plots and this would go on till we select the required number of sample units. This method of choosing a sample is known as sampling without replacement. This can be stated in different ways also. A set of 'n' sample units drawn from an aggregate of 'N' sample units is one of the $_NC_n$ possible sample sets. If our sampling process ensures that every sample has an equal chance of being chosen we will have obtained a random sample.

Sampling with and without replacement

Sampling without replacement : There are two ways of selecting 'n' units from the population of 'N' units, After drawing a sample unit from the population, it can be removed from the population before the next unit is drawn and this process can be continued till we get a sample of 'n' distinct units. Such a process is known as sampling without replacement.

Sampling with replacement : Alternatively, after drawing a sample unit from the population, it can be included in the population again, before the next unit is drawn. This process is continued till we get the required number of units. In this process a particular unit may be chosen more than once in which case the value of that particular unit should be considered as many times as it occurs in the sample. In a sample of 'n' units drawn with replacement there can be 'n' or less than 'n' distinct or different units.

Mean Values — Expectation and Uubiased Estimates

The meaning of these terms will be clear if we examine with reference to an illustration of a simple random sampling or probability sampling. When we take a random sample of 'n' units we can calculate any measure like Mean or Standard Deviation of the 'n' Values. These are called Sample Statistic. They are nothing but the estimates of the respective Mean and Standard deviation of the population of 'N' units.

However, if we select another sample of 'n' units, it may not give identical results.

Let us suppose that it is possible to construct all possible samples of 'n' units out of N population units. Let the sampling be without raplacement. The possible number of samples of 'n' will be equal to ${}_NC_n$ samples. When we select a sample of 'n' units we may have any one of the ${}_NC_n$ samples.

Let the sample be denoted by i and its sample mean be denoted by $\bar{x}_i$ Let the probability of selecting the sample be p_i. This can be arranged as follows :

Sl. No. of the sample	*Mean*	*Probability*
1	$\bar{x}_1$	p_1
2	$\bar{x}_2$	p_2
3	$\bar{x}_3$	p_3
...	...	...
i	$\bar{x}_i$	p_i
(i + 1)	$\bar{x}_{(i+1)}$	$p_{(i+1)}$
last ${}_NC_n$	$\bar{x}_{({}_NC_n)}$	$p_{({}_NC_n)}$

The expected overall mean of all sample means can be written as $\bar{x}$ which is equal to the population mean. This can be written as

$$E(\bar{x}) = \frac{\Sigma \bar{x}_i p_i}{\Sigma p_i}$$

Since $\Sigma p_i = 1$ (because the sum of the total probability is equal to 1) this can be written as

$$E(\bar{x}) = \sum_{i=1}^{{}_NC_n} \bar{x}_i p_i$$

If the expected value of the sample (statistic) is equal to the corresponding value of the population, the sample

statistic is said to be an unbiased estimate of the population parameter.

Example : Let us suppose we have a population of 5 units with their measurements as 50, 37, 26, 2, 0. The average of the five units is

$$\frac{50 + 37 + 26 + 2 + 0}{5} = \frac{115}{5} = 23$$

Instead of taking all the units, let us take a sample of two units (without replacement) and workout the mean value for all the samples of two units. There are ${}_5C_2$ samples.

$${}_5C_2 = \frac{\lfloor 5}{\lfloor 2 \times \lfloor 3} = \frac{5 \times 4}{1 \times 2} = 10 \text{ samples.}$$

The numbers of the 10 samples of two units are given below :

Sl. No.	Measures of the Samples (1)	(2)	Average $\bar{x}$	$\bar{x}^2$
1	50	37	43·5	1892·25
2	50	26	38·0	1444·00
3	50	2	26·0	676·00
4	50	0	25·0	625·00
5	37	26	31·5	992·25
6	37	2	19·5	380·25
7	37	0	18·5	342·25
8	26	2	14·0	196·00
9	26	0	13·0	169·00
10	2	0	1·0	1·00
Total			230·00	6718·00

Mean = **23·0**

It is seen that mean of all possible samples is equal to the mean of all the 5 units in the population. Hence the mean of a random sample is an unbiased estimate of the population Mean.

However, whether the sample 'statistic' is an unbiased estimate of the population parameter or not depends on the type of sampling procedure adopted and the statistics itself.

Random Sampling Errors (Slightly Advanced Portion)

It has been stated that a 'Statistic' say Mean ($\bar{x}$) calculated from a probability sample, i.e. a random sample (without replacement) of 'n' units from a population of 'N' units is only an unbiased estimate of population mean and hence it differs from the population parameter. Different samples of n' units may give dissimilar results of the Mean. But all these sample means cluster around a central value equal to E ($\bar{x}$) i.e. expected value of the Mean. This dissimilarity in the sample means occurs just because we take a random sample of only 'n' units instead of all the units in the population.

The extent of dissimilarity in these sample statistic is known as random sampling error which is generally measured by the standard deviation of the 'Statistic' from all possible samples. This is known as the Standard Error of Statistic. It is denoted by S. E. We can see from the above illustration

$$\text{S. E. of } \bar{x} = \sqrt{\frac{\Sigma \bar{x}^2}{n} - \bar{\bar{x}}^2} \text{ where}$$

$\bar{\bar{x}}$ is the overall average (In this example $\bar{\bar{x}} = 23$)

$$= \sqrt{\frac{6718}{10} - 529}$$

$$= \sqrt{\frac{6718 - 5290}{10}} = \sqrt{\frac{1428}{10}}$$

$$= \sqrt{142 \cdot 8} = 11 \cdot 95$$

We have seen from the above that the Mean of the sample means is same as the population mean. Let us also see for the sake of interest whether the standard deviation or standard error of the sample mean is equal to the standard deviation of the population. We have already calculated the Standard Deviation of the sample means equal to 11·95. Let us calculate the standard deviation of population.

	The value of population units x	x^2
	50	2500
	37	1369
	26	676
	2	4
	0	0
Total	115	4549

$$\bar{x} = \frac{115}{5} = 23. \qquad \bar{x}^2 = 529.$$

$$\sigma\bar{x} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

$$= \sqrt{\frac{4549}{5} - 529}$$

$$= \sqrt{\frac{4549 - 2645}{5}} = \sqrt{\frac{1904}{5}} = 19{\cdot}51$$

Let us tabulate the results of the population and the sample:

	Mean	Standard Deviation
Population	23	19·51
Sample	23	11·95

Though the Means are equal in both the cases, the standard deviations are not equal. We can find a strange

relationship between the standard deviation of the population and the standard deviation of the sample mean. Here the size of the sample is 2(n). We can find the following relationship is satisfied.

Standard Deviation of the sample Mean

$$= \frac{\text{Standard Deviation of the Population}}{\sqrt{\text{Sample Size} = n}}$$

$$= \frac{19{\cdot}51}{\sqrt{2}} = \frac{19{\cdot}51}{1{\cdot}41}$$

$$= 13{\cdot}9 = 14 \text{ approximately.}$$

The result we got earlier is *11.95* or *12* approximately and the difference is very appreciable. But generally the difference will not be appreciable. But the significant difference now noticed in this example is due to the fact that not only the size of the sample is small but the size of the population is also small.

$$\text{S. E. of } (\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

where σ : the standard deviation of the population

n : size of the sample.

Non-Sampling Errors

It has been already stated that we should adopt a very strict random method for selecting the samples. If any haphazard samples are selected, the result obtained may be biased. There is a chance of errors being crept in. These errors are called non-sampling errors or bias. There are several sources for such non-sampling errors and we should take all precautions to avoid them.

Sources of non-sampling errors

The possible sources of non-sampling errors are listed below :

(i) If the selection process deviates from the principles of probability sampling, bias is likely to arise.

(ii) Conscious or unconscious bias during the process of selecting a probability sample will lead to biased estimates. To eliminate error due to these sources it is always preferable to carry out the selection process in the central office by trained personnel and leave nothing or very little to the choice of the field investigator and that too with an adhoc rule of procedure.

(iii) An incomplete sampling frame introduces some error. The sample that we draw from such a frame will be from a portion of the population and hence it will lead to an estimate for that portion of the population only. If we accept this as the estimate for the entire population, there will be some error to an extent. Imperfections in sampling frame due to inaccuracies, duplication of units that are being out of date, introduce bias in the estimate.

(iv) Sometimes, it may be difficult to collect information from some of the units due to practical difficulties. For instance, in an enquiry where information are to be collected from some informants, he may not be available at his address when the investigator goes there or he may refuse to give any or part of the information. In an enquiry where information are collected by visiting fields, the investigator may find the fields flooded. Omission of randomly chosen units introduces an error in the estimate. Firstly, it will give an estimate for a portion of the population. Errors due to omissions of randomly chosen units, whatever may be the reason for omissions, are called the errors of non-response.

(v) The investigator, in his enthusiasm to return a certain volume of work, may substitute a convenient unit for a unit which he finds difficult to survey. This will lead to an error of the same nature as in (iv).

(vi) Inappropriate period of survey may lead to error in the estimate. It may not give estimate of the quantity that is being aimed at.

(vii) The questionnaire adopted for collecting the information may not contain questions to give all the required information. Answers to questions included may be difficult to obtain. Questions may lead to ambigious answers. The order in which questions are put and the way in which they are worded are also important. Inappropriate order, inappropriate arrangement and inappropriate wording of questions may introduce some error.

(viii) Method of collecting data is important. Different methods may suit different situations. For instance, the method of mailing questionnaire will not be suitable in India as the majority of the population is illiterate. If this method is adopted in India, there will be a good deal of non-response which introduces bias.

(ix) Faulty instructions and definitions of terms will introduce some error.

(x) Voluntary or involuntary errors in response can arise due to accidental mistakes in responding, failure of memory, bias due to lack of records, unwillingness to give the correct answer and so on.

(xi) Use of inaccurate and inappropriate instruments for measurement and methods of measurements may introduce bias.

(xii) The investigator can commit mistakes in recording, in understanding the questions and instructions and so on.

(xiii) Careless and disorganised field procedure may introduce some error.

(xiv) Errors can creep in during the processing stage and while interpreting the data, if appropriate statistical methods are not adopted.

The errors due to the above sources can be reduced to the minimum if these are kept in mind while planning and executing the survey. Pilot surveys will come to our aid in deciding the extent of error due to some of these sources. If the non-sampling errors due to various sources are of a cancelling nature, their net effect on the statistics, say $\bar{x}$ will be negligible. But we cannot always be sure of this.

The errors from different sources may all be one sided and the estimate calculated from the sample may be far from the true value. Whether the errors are of a cancelling nature or not, their effect on the estimated sampling error is considerable. These errors tend to increase the estimate of sampling error and hence the estimate of confidence interval. In order to accept the inference drawn from a sample, we should aim at reducing the estimated sampling error. Hence it is important to give sufficient attention to non-sampling errors at the time of planning and executing of survey.

Exercise

1. Define sample surveys. What are the advantages of sample surveys over complete enumeration?

2. Define :

   (i) Sampling Frame

   (ii) Standard Error

   (iii) Non-sampling Errors

   (iv) Random Number

   (v) Statistic

3. Define non-sampling errors and their sources.

4. Write an essay on non-sampling errors in Statistics.

CHAPTER III

THEORY OF SAMPLING

Principles of Sampling

The theory of sampling is based upon two important principles, namely

(1) The law of statistical regularity

(2) The law of inertia of large numbers.

Law of Statistical Regularity

Everything in nature and life occurs with a regularity. This is nothing but the reflection of Law of Statistical Regularity. This law of Statistical Regularity is the explanation for the fact that a sample duplicates an entire population in all its characteristics.

Law of inertia of large numbers

Though there may be changes in the characteristics of the individuals, the changes may not be appreciable when we consider the entire lot of the individuals. This is due to the fact that the difference between the individuals may be positive or negative. However, they may compensate when we consider the whole lot. The cumulative effect of the difference of each individual unit will not be there. This shows that in large numbers, changes would move slowly.

Sampling Distribution

We know that the collection of statistical information on census basis i.e. in respect of each and every individual in the universe is costly, time taking, besides laborious. Because of these difficulties statisticians resort to sample study. In this only a few units from the population are

selected and their behaviour studied and the result obtained from the sample is taken as the representative of the population.

Sample

A group of units or individuals or items selected from the population is called a Sample and each of the units in the sample may be called a sampling unit. The number of units in the sample is generally known as the size of the sample. Sometimes each of the units in the sample may itself be called a sample.

Random Sample

Selection of samples from the population is itself another branch of statistics called sampling technique. Generally, the samples selected with the help of random numbers to avoid personal bias of the enumerators and the investigators in the selection of the units are called Random samples and the process of selection is called Random Sampling.

Statistic

A statistical measure such as Mean or Standard Deviation computed from a sample is called a 'Statistic'. If we select a sample of required units from a population, we can calculate the Mean. If we select another sample consisting of the same number of units (and not the same members of units) we can also calculate the Mean for the second sample. Therefore, each of the Means calculated from each of the sample is itself an estimate of the Mean of the population. But each sample estimate of the Mean may differ from one another. This is because of the fact that the same units are not found in the different samples. In the same manner we can calculate as many number of means as there are samples of the same size.

Sampling Distribution of the 'Statistic'

Though each Mean, calculated from each of the sample is an estimate of population Mean, the mean of the means of the samples will be equal to the population Mean.

As we are having a frequency distribution for the units in the population we can also have a distribution for the statistic (Mean or standard deviation) calculated from each of the samples. Such a distribution of the statistic is called sampling distribution of the 'Statistic' As the average of the sampling Means is equal to the population Mean, we can normally expect the Mean or average of any 'statistic' to be equal to the value of the corresponding character of the population, as the size of the sample is increased and all possible samples are selected. Even if the original population is not normal, if large samples are taken the means of each such samples form a normal distribution with $\bar{x}$ as its mean and

$$\text{Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

where $\bar{x}$ = population Mean

σ = population S. D.

n = size of the sample.

Standard Error (S. E.)

It has been stated above that the Mean of any 'statistic' may normally be equal to the value of the corresponding character of the population. Though this is correct as far as Mean is concerned, it will not be exactly so in the case of a standard deviation. The Standard Deviation of Mean calculated from the sample or the sampling distribution is called the Standard Error (S. E.). The standard error of the statistics $\bar{x}$ will be equal to

$$\frac{\sigma}{\sqrt{n}}$$

Standard Deviation of the Mean

$$= \frac{\text{Standard Deviation of the Population}}{\sqrt{\text{Size of the sample}}}$$

$$\text{Standard Error or Standard Deviation of } \bar{x} = \frac{\sigma}{\sqrt{n}}$$

where σ = the standard deviation of the population.

n = the number of units in the sample or size of the sample.

The above formula can be easily understood from the following: When we calculate the Mean of the sample consisting of 'n' units, the differences noticed among the values of these 'n' units in the sample are disappearing from the Mean of the samples. The same situation happens in the case of each and every sample. Therefore the differences among the values of the Means of the different samples will not exhibit the same magnitude of difference noticed in the original values of the sample units or in the values of the population units. Thus the difference in the value of the Mean will be reduced to 1/nth of the difference noticed in the population units. Because, each of the sample Mean is calculated by dividing the value of the units in the sample by 'n' as the size of each sample is 'n'. In the same manner we can expect the value of the variance of the sample Mean will be 1/nth of the variance of the population unit or the variance of the sample since the latter is an estimate of the population value.

$$\text{Variance of the } \bar{x} = \frac{\text{Variance of the population}}{n}$$

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

Therefore the standard deviation of the Mean, otherwise known as standard error, S.E. will be equal to $\frac{\sigma}{\sqrt{n}}$ since the standard deviation is nothing but the square root of the variance.

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

$$S.D.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

If we want to reduce the value of the standard error, we have to increase the value 'n', i. e. the size of the sample. The value of the S. E. or the standard deviation of Mean is inversely proportional to the square root of the size of the sample.

Expectation of Sample Estimates

Whatever applies to the Mean will apply to the proportion also, since both of these are calculated with reference to the total value Σx. The properties of the random samples can be summarised as follows :

(i) The Mean of the Means derived from the sample approaches the population value (Population Mean) as the number of sample Means or the Means of the samples or indirectly the number of samples (not the number of units in the sample or size).

(ii) The Mean of the proportion of a particular character derived from the samples, approaches the population value, (population proportion) of that particular character as the number of samples increases.

(iii) When the number of samples is finite (when the number of samples is equal to the number of all possible samples of the same size from the population) the following conditions will prevail.

(a) the Mean of the sample Means will be equal to the population Mean and

(b) the Mean of the sample proportions will be equal to the population proportion.

In mathematical term, this property is expressed by the statement that the expected value of the Mean or the expected value of the proportion is equal to the population Mean or population proportion respectively. This statement is expressed in the following equation:

E (m) = u; E (p) = p; where 'E' means "Expectation of"; m = sample mean; u = population Mean; p = sample proportion; P = population proportion.

When the above condition or property does not hold good i. e. when the expectation of the (Mean) sample estimates is different from the population value, the sample estimate is said to be biased. Let us now see how the Standard Error of various parameters are derived.

1. Standard Error of the Sample Means

As the size of the sample increases, the Means of the different samples of the same size taken from the same population, though not equal to one another, will cluster more and more around the Mean of the population. If we calculate the Mean and the Standard Deviation for the sample means we find that the Mean of the Means will coincide with the population Mean while the standard deviation of the Mean decreases with the increase in the size of the sample.

The variability or variation among the Means of random samples of the same size is related to the variability or variation of the population by a definite mathematical formula namely,

Standard Deviation of the (m)

$$= \frac{\text{Standard Deviation of the population}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

where n = size of the sample.

Sampling Variance

The standard deviation of the Means of the samples is otherwise known as Standard Error. The square of the Standard Error or the variance of the Mean is called Sampling Variance of the Mean.

Confidence Limits

An important property of the random sampling is that the means of the random samples will be distributed normally

when the distribution of the units in the original population is either normal or approximately normal. This property of the random samples enables us to find with the help of the Normal Probability Integral Table, the limits within which any given proportion or the number of means of the samples or number of sample Means would lie.

We can say that 95% of the means of the samples of size 'n' would lie between the limits U—2 S. E and U + 2 S. E where U is the Mean of the population and $S.E = \frac{\sigma}{\sqrt{n}}$.

In other words, there is 95% chances for the means obtained from the samples to lie within this range. The probability of the sample means to lie within this ranges is 0·95.

Conversely if 'm' is the sample mean, the limits $m \pm 2$ S. E would contain the population mean (u) in 95 out of 100 cases. In such circumstances we may expect the following inequality to hold good on the average in 95% of the samples $m - 2\,S.E < U < m + 2\,S.E$. Therefore the probability for the inequality to hold good is 0·95. This is called confidence co-efficient and the range between the limits is known as confidence intervals. It can be noted that the range of the confidence interval will be smaller if the S.E. is smaller and the S. E will be smaller when the size of the sample is increased.

Standard Error of Sum of Means and Standard Error of Difference Means

Let us consider the Standard Error of (i) Sum of Means of two samples and (ii) Difference between Means of two samples. Let us take two samples from a population having a mean equal to $\bar{x}$ and standard deviation σ. The other particulars of the samples are as follows :

	Sample I	Sample II
Size of the sample	n_1 units	n_2 units
Mean of the sample	m_1	m_2
Difference of the Sample Mean	$m_1 - \bar{x}$	$m_2 - \bar{x}$
from the Population Mean	d_1	d_2

Sum of the Means $m_1 + m_2$

Difference of the Means $m_1 - m_2$

We know the variance of $m_1 = \frac{\sigma^2}{n_1}$

and the variance of $m_2 = \frac{\sigma^2}{n_2}$

where σ^2 is the variance of the population.

It has been proved that the variance of $(m_1 + m_2)$

$$= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

$$= \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$\therefore$ Standard Error of $(m_1 + m_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$\text{S. E. } (m_1 + m_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

It has also been proved S. E. $(m_1 - m_2)$

$$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Standard Error of Proportions

The Standard Error of an estimate of proportion can be given by the following formula :—

$$\text{S. E. } (P^1) = \sqrt{\frac{p(1-p)}{n}}$$

We can put $q = (1 - p)$

$$\therefore \text{S. E. } (P^1) = \sqrt{\frac{pq}{n}}$$

where p is the estimate of the proportion.

For calculating the S. E. from the sample, we must substitute for p, the value actually observed. The denominator can be slightly altered as (n — 1) instead of n. Hence the formula will also undergo a change.

$$\text{S. E. of (P)} = \sqrt{\frac{p(1-p)}{n-1}}$$

$$\text{and S. E. of } n_1 = n_1\sqrt{\frac{p(1-p)}{n-1}}$$

The Standard Error of the Sum or Difference of proportions from two samples

Let us consider the following samples :

	Sample I	Sample II
Proportion	p_1	p_2
Size of the sample	n_1	n_2

As in the case of mean, we can have the formula for S. E for the sum or difference of the population.

∴ Standard Error $(p_1 \pm p_2)$

$$= \sqrt{\frac{p_1(1-p_1)}{n_1-1} + \frac{p_2(1-p_2)}{n_2-1}}$$

Exercises

1. Write an essay on Theory of Sampling.
2. Define Standard Error and explain its uses in Test of Significance.
3. Explain the term '.Expectation of Sample Estimation'.
4. Write short notes on :

   (i) Statistic (ii) Standard Error (iii) Sampling variance (iv) Confidence limits

5. Find out from the following data the Sampling Error of the Mean

(i) Mean of the Sample 45 kg.

Standard deviation of the population	10 kg.
Size of the sample	25
(ii) Mean of the Sample	Rs. 70
Variance of the population	Rs. 12
Size of the sample	9

6. Calculate the Standard Error for the
   (i) sum of the sample means and
   (ii) difference of the sample means from the following data.

	Sample I	Sample II
Size of the sample	49	64
Mean	10 kg.	15 kg.

Standard Deviation of the population 10.

7. Calculate the Standard Error of the proportion

Size of the sample	25
Value of proportion in the population	1/4

8. Calculate the Standard Error for
   (i) the sum of the proportions
   (ii) the difference of the proportions in the samples.

	Sample I	Sample II
Size of the sample	49	64
Proportion	1/4	1/2

CHAPTER IV

TESTS OF SIGNIFICANCE

Tests of significance occupy an important place in the application of the statistical tool. Hence greater care has to be bestowed on this topic.

We have seen that the Mean of a sample taken from a population may not be exactly equal to the Mean of the population. Further, the mean of one sample taken from a population may not be equal to the mean of another sample taken from the same population. We also know that the measure to indicate the variability among the 'Statistic' calculated with the help of the sample is known as Standard Error (S.E.) of the character of parameter under study.

Inspite of the difference in the value of a sample Mean when compared with population Mean, we presume or assume that the sample is taken from that population. Similarly, inspite of the difference in the value of the Means of two samples, we may presume that the two samples belong to the same population. This means we are not attaching importance to the difference or rather we are ignoring the difference. In other words, we are assuming that the difference is not really a significant difference to consider the samples from different populations.

A question may arise. How far we can go on ignoring the difference in the values? Is there a limit to ignore the difference? These are very important questions to be considered in detail. In fact there is a limit to ignore the difference and beyond which importance has to be given to the difference and change our opinion also. There is a tolerance limit upto which we can allow the difference. If it exceeds the tolerance limit we may conclude (1) that

the sample is not from that population under question (2) that the two samples belong to two different populations. On the other hand, if the difference between means is less than the tolerance limit we may conclude that the sample or samples are from the same population.

Levels of Significance

There are two limits namely 5% level of significance and 1% level of significance. Generally 5% level significance would be sufficient. If we need greater accuracy, we should have 1% level of significance. As we are testing the significance of the difference rather than the difference, the process is called Testing of Significance and the limits are known as Level of Significance.

Normal Deviate

Generally the difference in the absolute values of the characteristic (Mean or Proportion or Variance or Correlation) as the case may be, is not directly considered for testing the significance. Instead, the difference is divided by the standard error of the particular characteristic (namely either Mean or Proportion or Variance or Correlation as the case may be) and converted into a ratio. This ratio is called the Normal deviate.

$$\left(\frac{x - \bar{x}}{\sigma}\right)$$

We can calculate from the Normal deviate table, the probability corresponding to the Normal deviate of the difference calculated. We should then verify whether the probability obtained for the normal deviate is either less than 0.05 or less than 0.01 as the case may be.

Interpretation

In case we are adopting a 5% level of significance and the value of the probability for the normal deviate of the difference is less than 0.05, then we may say that the probability for a difference of the given order or magnitude is less than 0.05. Hence the difference is not due to any chance

or sampling fluctuation, and hence the difference is really a significant difference. In the same manner we can test the significance at 1% level. If the probability for the normal deviate of the difference is less than 0.01, then we may conclude that the difference of the order noticed between the values is not due to chance or sampling fluctuation, and hence the difference is really significant. On the other hand, if the probability obtained for the normal deviate of the difference is greater than the required level of probability 0.05 or 0.01, we may conclude that there is really a greater chance to have a difference of the observed magnitude and hence the difference cannot be said to be significant or the difference is said to be not significant.

Errors of Judgement

In this process, of course, there is one danger of committing an error. A significant difference may be decided as non-significant and non-significant difference may be declared as significant difference. We also indirectly accept this in view of the level of significance adopted. Great controvercy is still going on among statisticians about the safety of this application because of the two kinds of errors enumerated above. Still this test is widely used in all statistical investigations.

Test of significance in practice

In actual practice, testing is not done on the basis of comparing the probability of the computed normal deviate of the difference with either 0.05 or 0.01 probability depending upon the level of significance required. Instead, the computed normal deviate of the difference itself is compared with the normal deviate either for 0.05 probability or 0.01 probability. The normal deviate corresponding to 0.05 probability is 1.96 or 2 approximately and the normal deviate for 0.01 probability is 2.58 or approximately 3. The advantage is we need not refer to the normal deviate table every time since 1·96 and 2.58 are constants for 0.05 and 0.01 probability respectively.

If the computed value of the normal deviate of the difference is greater than 1.96, it is said that the difference

noticed in the value is really significant at 5% level. On the other hand if the computed value of the normal deviate is less than 1.96 it will be said that the difference is not significant. The same type of argument will be advanced if the computed value is greater or less than 2·58 for 1% level of significance.

Null Hypothesis

It may be noted that in all cases we proceed from the assumption (a) that the sample is taken from the population; (b) the two different samples are taken from the same population. Indirectly it means that there is no difference (i) between the sample value and the population value of the parameter (ii) between the parameters of the different samples. This type of assumption or hypothesis is called Null Hypothesis since the basic principle is that the difference noticed between the values is 'Null' or 'Nil' or '0'.

After assuming the Null Hypothesis we proceed further to test the validity of this assumption on the basis of the details available in the problem. Either we may reject the null hypothesis or accept the null hypothesis. Rejection of the hypothesis indicates the presence of significant difference and the acceptance of the hypothesis indicates the difference present as insignificant.

Application of the tests

We shall confine our study of the test of significance to test the difference noticed,

(a) between the Mean of the sample and the Mean of the population.

(b) between the Means of two different samples.

(c) between the proportions of a particular characteristic of the sample and the population.

(d) between the proportions of a particular characteristics of two different samples.

Computation of Normal deviate

In all these tests we proceed from the Normal deviate. In order to compute the normal deviate of any characteristic, we should know the standard error of the characteristics (Mean or proportion as the case may be). If we want to calculate the S. E. of the characteristics, we should have the standard deviation of the population. In certain cases the standard deviation or the variance of the population may not be available and they have to be estimated from the sample itself.

Types of sample

There are two types of samples namely, large sample and small sample. If the size of the sample or the number of units in the sample is 30 and above it is called a large sample and others are called small samples. The method of estimation of the population variance from large sample is different from the method adopted in the case of small samples.

1. Testing the significance of the difference between Sample Mean and the Population Mean

1. In this case we shall first consider a large sample consisting of more than 30 units.

Example 1

A popular tyre company had advertised that its products are highly reliable saying that its tyres would run an average distance of 16000 km without any necessity for retreading. Its standard deviation is given as 1500 km per tyre. A lot of 100 tyres, were purchased and the average running life of these tyres is 15500 km. Can we say whether these 100 tyres are products of the above Company ?

Population mean = 16000 km.
Population standard deviation = 1500 km.
Size of the sample = 100 tyres
Mean of the sample = 15500 km.

These are the data available in this problem. Let us assume that the sample belongs to the same company and thereby we presume that there is no difference between the two means. But actually there is difference between the Sample Mean and Population Mean.

The difference between the Means = 16000 — 15500

= 500 (We have not considered the sign.)

S. D. of the Sample Mean or S. E. of the Sample Mean

$$= \frac{\text{Standard deviation of the population}}{\sqrt{\text{Size of the sample}}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{1500}{\sqrt{100}} = \frac{1500}{10} = 150 \text{ km.}$$

The normal deviate corresponding to

the difference in the Mean $= \dfrac{X - \bar{X}}{\sigma X}$

$$= \frac{500}{150} = \frac{10}{3} = 3{\cdot}3$$

The value 3·3 is greater than 1·96, the normal deviate at 5% level of significance and it is also greater than 2·58, the normal deviate at 1% level of significance.

Hence the difference between sample mean and population mean is significant and so we reject Null Hypothesis. Therefore, the lot does not belong to the product of the company which had advertised and it belongs to some other company.

Example 2

We have purchased another lot of 36 tyres from another dealer. The average length of its running is 16600 km.

Can this lot belong to the above Company which has advertised its products saying its average life is 16000 km with a standard deviation of 1500 km per tyre.

Population Mean = 16000 km.

Standard Deviation of the population = 1500 km.

Sample Mean = 16600 km.

Size of the sample = 36 = n.

Difference between the sample mean and the population mean

$$= 16000 - 16600$$

$$= 600 \text{ (sign need not be considered.)}$$

Standard deviation or

$$\text{S. E. of the Sample Mean} = \frac{\text{S. D. of the population}}{\sqrt{\text{Size of the sample}}}$$

$$= \frac{1500}{\sqrt{36}} = \frac{1500}{6} = 250 \text{ kms.}$$

Normal deviate corresponding to the difference of the mean

$$= \frac{\text{Difference}}{\text{S. E.}}$$

$$= \frac{600}{250} = 2{\cdot}40$$

The computed value of the Normal deviate is greater than the normal deviate at 5% level of significance (1·96) and less than the normal deviate at 1% level of significance (2·58). Hence the difference between the population mean and sample mean is significant at 5% level and not significant at 1% level.

Therefore, we have to reject the null hypothesis at 5% level and accept it at 1% level. When we consider 5% level

of significance we can say that the product does not belong to the same company. At 1% level, the product can be said to belong to the same company.

A. Testing the significance of the difference between the Means of two samples

Example 3

A certain intelligent test was applied to a large group of students and found that the S. D. of the group is 40 score. The test is given to another group of 36 boys and found that the average score is 150. The test is given to another group of 64 boys and the average score is 160. Does it show any significant difference ?

	Sample I	Sample II
Size	$n_1 = 36$	$n_2 = 64$
Mean	$m_1 = 150$	$m_2 = 160$

Standard Deviation of the population = 40.

S.E. of the difference of the Means

$$S.E.(m_1 - m_2) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 40\sqrt{\frac{1}{36} + \frac{1}{64}}$$

$$= 40\sqrt{\frac{100}{36 \times 64}}$$

$$= \frac{40 \times 10}{6 \times 8} = 8{\cdot}33$$

Difference between the Means = 160 — 150 = 10.

Ratio similar to Normal Deviate for the corresponding difference of the Means $= \frac{10}{8{\cdot}33}$ which is equal to 1·2. Hence the difference is not significant both at 5% and 1% levels since the ratio is less than 1·96 and 2·58. So the two

samples belong to the same population with Standard Deviation = 40.

Let us consider the following case :

	Sample I	Sample II
Size of the sample	$n_1 = 81$	$n_2 = 100$
Mean	$m_1 = 6{\cdot}0$	$m_2 = 4{\cdot}5$
Variance	$v_1 = 10{\cdot}5$	$v_2 = 10{\cdot}25$

In this problem as the variance of the population is not given, it has to be determined from the sample.

Difference in the Mean $= m_1 - m_2 = 6{\cdot}00 - 4{\cdot}50 = 1{\cdot}50$

$$S.\ E\ (m_1 - m_2) = \sqrt{\frac{v_1}{n_1} + \frac{v_2}{n_2}}$$

$$= \sqrt{\frac{10{\cdot}50}{81} + \frac{10{\cdot}25}{100}} = 0{\cdot}48$$

$$\text{Ratio} = \frac{\text{Difference in the Mean}}{\text{S. E. of the difference of the Mean}}$$

$$= \frac{1{\cdot}5}{0{\cdot}48} = 3{\cdot}12$$

This shows that the difference is significant at 5% level and at 1% level.

B. Testing the significance of the proportion

(i) Testing the difference between the sample proportion and the observed proportion:

Example 1

A coin was tossed 100 times. The head turned on 65 occasions. Examine whether the coin is good.

Let us assume that the coin is good. If the coin is good, the head should turn up on 50 occasions while the tail

on 50 occasions. This will be the position in the population. Therefore, the proportion of the head to turn is

$$\frac{50}{100} = 0{\cdot}5$$

The proportion in the population $= 0{\cdot}5$

The proportion in the sample $= \frac{65}{100} = 0{\cdot}65$

Difference between the two proportions
(Sample proportion and population proportion) $= 0{\cdot}50 - 0{\cdot}65$
$= 0{\cdot}15$ (sign is not considered)

Standard Error of the proportion $= \sqrt{\frac{p\,q}{n}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{100}}$

$$= \sqrt{\tfrac{1}{2} \times \tfrac{1}{2} \times \frac{1}{100}} = \frac{1}{20} = 0{\cdot}05$$

The ratio similar to normal deviate for the corresponding value of the difference in the proportion is

$$\frac{\text{Difference in the proportion}}{\text{Standard Error of the proportion}}$$

$$= \frac{0{\cdot}15}{0{\cdot}05} = 3.$$

Since the value 3 is greater than 1·96 and 2·58, we say the difference noticed at both the levels is significant. Hence the assumption is rejected. Therefore, the coin is not good or the coin is biased at both 1% and 5% levels.

(ii) Testing the significance of the difference between the proportions of two samples :

Example 2

During the country-wide investigation, the incidence of a particular disease was found to be 2%. In a college with a strength of 500 students, 5 were reported to be affected by the same disease while in another college, with a strength of 1500 students, 30 were affected. Does this indicate any significant difference?

	Sample I	Sample II
Number of students	$n_1 = 500$	$n_2 = 1500$
No. of students affected	$x_1 = 5$	$x_2 = 30$
Proportion	$p_1 = \frac{5}{500}$	$p_2 = \frac{30}{1500}$
	$= 0{\cdot}01$	$= 0{\cdot}02$

Difference between the sample proportions $0{\cdot}02 - 0{\cdot}01 = 0{\cdot}01$
Population proportion $P = 0{\cdot}02$.

Standard Error of the difference of the proportion:

$$S.E(P_1 - P_2) = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$P = 0{\cdot}02 \text{ and } q = 1 - P = 1 - 0{\cdot}02 = 0{\cdot}98$$

$$S.E(P_1 - P_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{0{\cdot}02 \times 0{\cdot}98\left(\frac{1}{500} + \frac{1}{1500}\right)}$$

$$= \sqrt{0{\cdot}02 \times {\cdot}98 \times \frac{4}{1500}}$$

$$= \sqrt{\frac{0{\cdot}0196}{375}}$$

$$= 0{\cdot}0073$$

The ratio similar to the normal deviate corresponding to the difference between the proportion:

$$\frac{\text{Difference in the proportion}}{\text{S. E. of the proportion}} = \frac{0{\cdot}01}{0{\cdot}0073}$$

$$= 1{\cdot}4$$

The value is less than $1{\cdot}96$ and $2{\cdot}58$. Hence the difference is not significant at both 5% and 1% levels of significance.

It may be noted that the proportion of the population is given in the problem. But in certain cases the proportion of the

population will not be available and we have to estimate the population proportion from the sample proportion itself. Let us take the same population without the population proportion given.

Population proportion

$$= \frac{n_1 p_1 + n_1 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{5 + 30}{500 + 1500} = \frac{35}{2000}$$

$$= 0{\cdot}0175.$$

$$q = 1 - p = 1 - 0{\cdot}0175 = 0{\cdot}9825.$$

The S. E. of the difference of the proportion

$$= \sqrt{0{\cdot}0175 \times 0{\cdot}9825 \times \frac{4}{1500}} = 0{\cdot}007$$

The ratio corresponding to the normal deviate

$$= \frac{0{\cdot}01}{0{\cdot}007} = \frac{10}{7} = 1{\cdot}4$$

The difference is not significant since the computed value is less than 1·96 and 2·58.

Exercise

1. Explain the Test of Significance in Statistical Analysis and bring out its uses in statistical application.

2. Explain the Level of Significance.

3. Write short notes on

   (i) Level of Significance

   (ii) Null Hypothesis

   (iii) Normal Deviate.

4. A sample of 1000 members has a mean of 45 kg with a Standard Deviation of 9 kg. Test whether the sample is from a population with Mean 4 kg and the Standard Deviation 9 kg.

5. Test whether the following two samples are from the sample population with Standard Deviation 15 kg.

	I	II
Size of the sample	25	81
Mean	10	12

6. In a sample of 500 men from a city, 400 are found to be smokers. In another sample of 1000, the smokers are found to be 800. Do they indicate any significant difference ?

7. A coin is tossed 350 times and it is found that head occurred 150 times. Test whether the coin is biased.

CHAPTER V

ASSOCIATION OF ATTRIBUTES

Statistics of Attributes

We have already studied that data can be collected on qualitative as well as quantitative characteristics. If the population is divided on the basis of sex, literacy, employment, it is said to be qualitative classification. On the other hand, if the population is classified according to the size of the income it is said to be quantitative classification. The observations based on descriptive characteristics are termed as Statistics of Attributes. So far, we have studied about the statistics of variables. We shall now see the relationship between two attributes and how it can be established by the method of Association.

Classes and their frequencies

In this process, we can say whether a particular unit has a particular characteristics or not. In other words, a particular characteristics may be present or absent in a particular unit. The general practice is that the presence of the characteristics is represented by the capital letters like A, B, C etc. and their absence will be represented by the corresponding Greek Letters α, β, γ, etc.

The individuals possessing the attribute A, are said to belong to Class A. The number of individuals belonging to Class A is called the frequency of class A. The frequency of class A is represented by the class within brackets (A). In the same way, the frequencies of classes B and C will be written as (B) and (C) respectively. The individuals who do not possess the attribute A are said to belong to the class α. The number of individuals who do not possess that attribute A is called the frequency of class α and is denoted by (α).

The possession of two attributes is denoted by the letters placed side by side within the brackets as (AB) or $(\alpha\beta)$ or $(\beta\alpha)$ or $(A\beta)$.

It can be explained as follows:

(AB) : The No. of individuals with possession of A and B.

$(A\beta)$: The No. of individuals with possession of A and with possession of β (possession of A and the No. of individuals with absence of B).

$(B\alpha)$: The No. of individuals with possession of B and possession of α (possession of B and absence of A).

$(\alpha\beta)$: The No. of individuals with possession of α and possession of β (the absenee of A and B).

Positive and Negative Attributes

The attributes denoted by the capitals A, B, C etc. may be termed as the positive attributes. Their contraries denoted by the letters α. β, γ are called Negative attributes. The classes A, B, C are called positive classes while α, β, γ are called negative classes. The following classes are called pair of contrary classes.

AB and $\alpha\beta$

$A\beta$ and $B\alpha$

Order of classes

A class possessing one attribute is known as the class of first order. A class possessing two attributes is known as a class of second order. Thus the classes A, AB are called first and second order classes respectively. Similarly the class frequencies (A), (B) are called the first order frequencies and the frequencies $(A\beta)$, $(B\alpha)$, (AB), $(\alpha\beta)$ are called the second order frequencies.

When no attributes are specified, the total number of observations constitutes the universe with its limits specified and it will be denoted by the letter N.

Ultimate classes and Ultimate class frequencies

The classes specified by the highest order are termed as ultimate classes, and consequently their frequencies are termed as ultimate class frequencies. If we know the frequencies of classes AB and Aβ, i.e. (AB) and (Aβ), we can find the (A). Similarly, if we know the frequencies of classes $\alpha\beta$ and αB i.e. the frequencies ($\alpha\beta$) and (αB) we can find (α). Once we find (A) and (α) we can find N since (A) + (α) = N. This is due to the fact that the total frequencies can be divided into 2 parts namely, (1) those possessing the attributes and (2) not possessing the attributes.

In this manner we can establish the following facts:

$$(A) + (\alpha) = (N) \quad \text{......(1)}$$

$$(B) + (\beta) = (N) \quad \text{......(2)}$$

$$(AB) + (A\beta) = (A) \quad \text{......(3)}$$

$$(AB) + (\alpha B) = (B) \quad \text{......(4)}$$

$$(\alpha\beta) + (\alpha B) = (\alpha) \quad \text{......(5)}$$

$$(\alpha\beta) + (A\beta) = (\beta) \quad \text{......(6)}$$

From (1) and (2) we can have

$$(A) + (\alpha) = (B) + (\beta) = N \quad \text{......(7)}$$

From (3) and (5) we can have

$$(AB) + (A\beta) + (\alpha\beta) + (\alpha B) = N \quad \text{......(8)}$$

From (4) and (6) we can have

$$(AB) + (B\alpha) + (\alpha\beta) + (A\beta) = N \quad \text{......(9)}$$

We know (8) = (9)

Example

From the following ultimate frequencies, find the frequencies of the positive and negative classes.

(AB) = 125; (αB) = 50; ($\alpha\beta$) = 75 and (Aβ) = 60.

Let us first calculate 'N'.

$$N = (AB) + (\alpha B) + (\alpha\beta) + (A\beta)$$
$$= 125 + 50 + 75 + 60 = 310.$$

Positive cases are A, B and AB

∴ The positive frequencies are (A), (B), (AB).

Similarly negative classes are α, β and α β and

the negative frequencies (α), (β) and $(\alpha\beta)$.

We know that

$$(A) = (AB) + (A\beta) = 125 + 60 = 185$$

$$(B) = (AB) + (\alpha B) = 125 + 50 = 175$$

$$(\alpha) = (\alpha\beta) + (B\alpha) = 75 + 50 = 125$$

$$(\beta) = (\alpha\beta) + (A\beta) = 75 + 60 = 135.$$

The value of (α) and (β) can be indirectly calculated from the value of (N), (A) and (B).

$$N = (A) + (\alpha)$$

$$310 = 185 + (\alpha)$$

$$\therefore (\alpha) = N - (A)$$

$$= 310 - 185$$

$$= 125.$$

Similarly

$$N = (B) + (\beta)$$

$$310 = 175 + (\beta)$$

$$\therefore (\beta) = 310 - 175$$

$$= 135.$$

The above details can be represented in the following table:

(AB)	(Bα)	B
(Aβ)	(αβ)	β
A	(α)	N

125	50	175
60	75	135
185	125	310

Once we construct the table and fill them with their respective frequencies we can find the value of (A), (B) and N.

Independence of Attributes

If there is no relationship between two attributes then they are said to be independent. If two attributes, say A and B are independent then the following condition will be satisfied.

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(A)}{N}$$

1. $\frac{(AB)}{(B)}$ = Proportion of A's among B's.

2. $\frac{(A\beta)}{(\beta)}$ = Proportion of A's among β's (or) Proportion of A's among non B's

3. $\frac{(A)}{N}$ = Proportion of A's among the whole group.

Similarly we can have

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} = \frac{(B)}{N}$$

1. $\frac{(AB)}{A}$ = Proportion of B's among A's

2. $\frac{(\alpha B)}{\alpha}$ = Proportion of B's among α's (or) Proportion of B's among non A's

3. $\frac{(B)}{N}$ = Proportion of B in the whole group

$$\therefore \frac{(AB)}{(A)} = \frac{(B)}{N}$$

$$(AB) = \frac{(A)(B)}{N}$$

If we prove that $(AB) = \frac{(A)\ (B)}{N}$ then we say that the two attributes A and B are independent.

When the two attributes A and B are independents, then their contraries namely the attributes α and β are also independent. If α and β are independent then the following condition will be satisfied.

$$(\alpha\ \beta) = \frac{(\alpha)\ (\beta)}{N}$$

Example 1

If (A) = People inoculated = *100*

(B) = People not attacked by fever = *120*

(AB) = People inoculated and not attacked = *40*

N = Total number of people = *300*

find out whether A and B are independent. For this we have to find out whether the following condition holds good.

$$(A\ B) = \frac{(A)\ (B)}{N}$$

$$40 = \frac{100 \times 120}{300}$$

$$= 40$$

It holds good.

Hence A and B are independent. So inoculation and immunity from fever are independent.

Example 2

In the above example $(\alpha\beta)$ is given *120*. Find out whether α and β are independent.

Before finding out whether α and β are independents, we have to first find out (α) and (β).

We know

$$(A) + (\alpha) = N.$$

$$\therefore \alpha = N - (A)$$

$$= 300 - 100$$

$$= 200.$$

Similarly we can find out the value of (β).

$$(B) + (\beta) = N$$

$$\therefore \quad (B) = N - (B)$$

$$= 300 - 120$$

$$= 180.$$

Let us now find out whether

$$(\alpha\beta) = \frac{(\alpha)\ (\beta)}{N} = \frac{200 \times 180}{300} = 120.$$

Hence α and β are independents. i. e. Attack of fever and non inoculation are independent of each other.

CONTINGENCY TABLE

When we are given two attributes A and B, the positive and negative, the ultimate class frequencies of these attributes can be presented in the form of a table as given on the next page. This table is called contingency table.

Attributes	A	α	Total
B	(AB)	(Bα)	(B)
β	(Aβ)	($\alpha\beta$)	(β)
Total	(A)	(α)	N

It may be seen that

1. $(A) = (AB) + (A\beta)$
2. $(\alpha) = (B\alpha) + (\alpha\beta)$
3. $(B) = (AB) + (B\alpha)$
4. $(\beta) = (A\beta) + (\alpha\beta)$
5. $N = (A) + (\alpha) = (B) + (\beta)$
   $= (AB) + (A\beta) + (B\alpha) + (\alpha\beta)$

From the above relationship we can find out the missing frequencies and any other values.

Example

Find out the missing frequencies from the following data.
$(A) = 185$; $(B) = 175$; $(AB) = 125$; $N = 310$.

The missing frequencies are (α), (β), $(A\beta)$, $(B\alpha)$, and $(\alpha\beta)$.
We know that

(i) $(A) + (\alpha) = N$

$$\therefore (\alpha) = N - (A)$$

$$= 310 - 185$$

$$= 125.$$

(ii) $(B) + (\beta) = N$

$$\therefore (\beta) = N - (B)$$

$$= 310 - 175$$

$$= 135.$$

(iii) $(AB) + (A\beta) = (A)$

$$\therefore (A\beta) = (A) - (AB)$$

$$= 185 - 125$$

$$= 60.$$

(iv) $(AB) + (B\alpha) = (B)$

$$\therefore (B\alpha) = (B) - (AB)$$

$$= 175 - 125$$

$$= 50.$$

With the details now computed we can construct the following contingency table.

	A	α	Total
B	(AB) 125	($B\alpha$) 50	(B) 175
β	($A\beta$) 60	($\alpha\beta$) 75	(β) 135
Total	(A) 185	(α) 125	N 310

ASSOCIATION AND DISASSOCIATION

We have said that two attributes A and B are independent when $(AB) = \frac{(A)\ (B)}{N}$

If (AB) is not equal to $\frac{(A)\ (B)}{N}$ then A and B are not independents. In other words A and B are associated. (AB) may be greater than $\frac{(A)\ (B)}{N}$ or (AB) may be less than $\frac{(A)\ (B)}{N}$

The association may be either positive or negative.

$$\text{If } (AB) > \frac{(A)\ (B)}{N}$$

i. e. if $(AB) - \frac{(A)\ (B)}{N}$ is equal to a positive quantity, A and B can be said to be positively associated or simply associated. If $(AB) < \frac{(A)\ (B)}{N}$ or $(AB) - \frac{(A)\ (B)}{N}$ is equal to a negative quantity, A and B are said to be negatively associated or disassociated. Hence the value $(AB) - \frac{(A)\ (B)}{N}$ can be taken as the indicator.

1. If $(AB) - \frac{(A)\ (B)}{N} = 0$ then A and B are independents.
2. If $(AB) - \frac{(A)\ (B)}{N}$ = a positive quantity, then A and B are associated.
3. If $(AB) - \frac{(A)\ (B)}{N}$ = a negative quantity, then A and B are disassociated.

However, the expression

$$(AB) - \frac{(A)\ (B)}{N} \text{ can be simplified as}$$

$$\frac{1}{N}\left\{ (AB)\,N - (A)\,(B) \right\}$$

$$= \frac{1}{N}\left\{ (AB)\,[(AB) + (A\beta) + (B\alpha) + (\alpha\beta)] \right.$$

$$\left. - \left\{\left[(AB) + (A\beta)\right]\left[(AB) + (B\alpha)\right]\right\}\right\}$$

$$= \frac{1}{N}\left\{ (AB)\,(\alpha\beta) - (A\beta)\,(B\alpha) \right\}$$

Therefore, if

(i) $(AB)\ (\alpha\beta) - (A\beta)\ (B\alpha) = 0$
A and B are independent.

(ii) $(AB)\ (\alpha\beta) - (A\beta)\ (B\alpha) = +$ ve
A and B are associated.

(iii) $(AB)\ (\alpha\beta) - (A\beta)\ (B\alpha) = -$ ve
A and B are disassociated.

It may be seen that $(AB)\ (\alpha\beta)$ is the product of the frequencies in the diagonal classes.

Similarly $(A\beta)\ (B\alpha)$ is the product of the frequencies in the diagonal classes of the contingency table.

We know that

$(AB)\ (\alpha\beta)$ is the product of pair of contrary classes. Similarly $(A\beta)\ (B\alpha)$ is also the product of the pair of contrary classes.

Therefore the independence or association or disassociation of two attributes can be determined by the difference between the product of the two pairs of contrary classes.

Co-efficient of Association

In order to compare the degree of association between two attributes A, B in the two groups, Yule has given the following co-efficient of association. It is denoted by Q.

$$Q = \frac{(AB)\ (\alpha\beta) - (A\beta)\ (B\alpha)}{(AB)\ (\alpha\beta) + (A\beta)\ (B\alpha)}$$

$$= \frac{\text{Difference of the product of the pairs of contrary classes}}{\text{Sum of the product of the two pairs of contrary classes}}$$

If $Q = 0$, A and B are independent.

i. e. If $(AB)\ (\alpha\beta) - (A\beta)\ (B\alpha)$ is equal to 0, A and B are independent.

If Q is a positive, A and B are associated.

If Q is negative, A and B are disassociated.

Example :

Calculate the co-efficient of association for the following data :

	A	α	Total
B	80	10	90
β	40	20	60
Total	120	30	150

$$Q = \frac{(AB)\ (\alpha\beta) - (A\beta)\ (B\alpha)}{(AB)\ (\alpha\beta) + (A\beta)\ (B\alpha)}$$

$$= \frac{80 \times 20 - 10 \times 40}{80 \times 20 + 10 \times 40}$$

$$= \frac{1600 - 400}{1600 + 400} = \frac{1200}{2000} = 0{\cdot}6$$

It indicates a positive association.

Difference between Association of Attributes and Correlation

Both correlation and association of attributes are important statistical tools to study the relationship between variables. When the given variables are quantitative varibles the relationship can be studied with the help of correlation. If the variables given are qualitative variables the relationship can be studied with the help of association of attributes.

Exercise

1. Define co-efficient of Association

   In an experiment on immunization of cattle from tuberculosis, the following results were obtained.

	Died	Unaffected
Inoculated	12	26
Not inoculated	16	6

   Examine the effect of Inoculation in controlling the susceptibility to tuberculosis.

2. Investigate the association between the eye colour in Mother and daughter from the following data:

   Both Mothers and daughters with dark eyes = 200

   Mothers without dark eyes and daughters with dark eyes = 360

   Mothers with dark eyes and daughters without dark eyes = 320

   Both Mothers and daughters without dark eyes = 120

3. Find whether the data given below are consistent.

   $A = 25, \quad B = 20$

   $AB = 10, \quad N = 30$

4. The following data are given. Find out whether attributes A and B are independent.

   $A = 30, \quad B = 6, \quad AB = 12, \quad N = 150.$

CHAPTER VI

ANALYSIS OF VARIANCE AND DESIGN OF EXPERIMENTS

We have already studied about the variance and the standard deviation as measures of dispersion which can be used for comparing different distributions. In this chapter we shall study further about the application of these measures and more particularly about the variance.

It is a well known fact that all the units either in a population or in a sample may not have equal values and difference among their values is inevitable. We have measures namely the variance and standard deviation to measure the average difference in the value per head or per unit.

There may be different factors responsible for the difference in the values of the different items. Therefore, it has become necessary to find out the contribution of each such factor for the difference noticed in the values. Further it is also necessary to find out whether the difference in the value contributed by each such factor is really significant or whether such difference in the values is quite likely in the normal course. A detailed study of the problem of this kind is called Analysis of Variance since the total variance found is analysed according to different factors of contribution.

We shall study about this in detail with the help of an illustration. The following is the yield of paddy (kg per plot) obtained from crop cutting experiments conducted in 12 plots of uniform size in respect of four varieties in three districts.

Districts	Varieties			
	V_1	V_2	V_3	V_4
D_1	20	25	22	21
D_2	23	26	25	22
D_3	26	27	28	23

These details can further be simplified as follows:

Districts	Varieties				District Total	District Average in kg.
	V_1	V_2	V_3	V_4		
D_1	20	25	22	21	88	22
D_2	23	26	25	22	96	24
D_3	26	27	28	23	104	26
Variety Total	69	78	75	66	288	
Variety Average	23	26	25	22		24 kg per plot

There are three districts and four varieties of paddy. Each variety was harvested from each of the three districts and hence there are 3 types of yield for each variety of crop. Similarly for each district there are 4 different yields also. Let the three districts be denoted by the letter D_1, D_2 and D_3 and the four varieties be denoted by V_1, V_2, V_3 and V_4 respectively.

In the initial stage, let us ignore the existence of the districts and varieties and consider the yields of all the 12 experiments as a single sample from the State. We shall calculate the variance per plot for this sample in the usual manner.

Yield per plot x	x^2
20	400
23	529
26	676
25	625
26	676
27	729
22	484
25	625
28	784
21	441
22	484
23	529
Total 288	6982

$$\text{Mean } \bar{x} = \frac{288}{12} = 24$$

$$\bar{x}^2 = 24 \times 24 = 576.$$

$$\text{Variance (V)} = \frac{\Sigma x^2}{N} - \bar{x}^2$$

$$= \frac{6982}{12} - 576$$

$$= 581\frac{10}{12} - 576$$

$$= 5\frac{10}{12} = 5\frac{5}{6} \text{ kg per plot.}$$

This is the average square of the difference or Mean Square difference per plot in the sample. The average difference per plot will be equal to

$$\sigma = \sqrt{5\frac{5}{6}} = \sqrt{\frac{35}{6}} = 2{\cdot}4 \text{ kg per plot.}$$

Analysis : District Approach (Ignoring the presence of varieties)

Let us now split the variance. For this purpose we shall consider only the three districts. Since there are three different districts, the variance between the districts may be a factor responsible to some extent for the difference which is expressed in terms of variance before.

In order to compare and compute the variations between the districts we shall consider the district average yields. The district average yields are 22, 24 and 26 kgs and the State average is 24 kg. As usual, we shall compute the variance for the district average.

District average

	$\bar{x}$	$\bar{x}^2$
	22	484
	24	576
	26	676
Total	72	1736

Average: $\bar{\bar{x}} = 24$; $\bar{\bar{x}}^2 = 24 \times 24 = 576$.

$$V = \frac{\Sigma \bar{x}^2}{N} - \bar{\bar{x}}^2$$

$$= \frac{1736}{3} - 576$$

$$= 578\frac{2}{3} - 576 = 2\frac{2}{3} \text{ kg per district.}$$

Even though, there are variations between the districts, we find that the yield rate within a particular district is not uniform. This shows that there are different kinds of variations within the districts. Since there are 3 districts, the variation within the district itself may be sub-divided into three portions. Broadly speaking, the variance found in the sample can be analysed into two parts namely (1) the variation between the districts and (2) the variation within the district.

The variance within each of the three districts can be computed as follows by comparing the district yield with respective district average.

	Yield District I		Yield District II		Yield District III	
	x	x^2	x	x^2	x	x^2
	20	400	23	529	26	676
	25	625	26	676	27	729
	22	484	25	625	28	784
	21	441	22	484	23	529
Total	88	1950	96	2314	104	2718
$\bar{x}$ =	22		24		26	
$\bar{x}^2$ =	484		576		676	

Variance $= \frac{1950}{4} - 484 \qquad \frac{2314}{4} - 576 \qquad \frac{2718}{4} - 676$

$= 487\frac{2}{4} - 484 \qquad 578\frac{2}{4} - 576 \qquad 679\frac{2}{4} - 676$

$= 3\frac{1}{2} \qquad 2\frac{1}{2} \qquad 3\frac{1}{2}$

Total of 3 districts $= 3\frac{1}{2} + 2\frac{1}{2} + 3\frac{1}{2} = 9\frac{1}{2}$

$\therefore$ Average per district $= 9\frac{1}{2} \div 3 = \frac{19}{2} \times \frac{1}{3}$

$= \frac{19}{6} = 3\frac{1}{6}$

We have calculated the following:

1. The variance between the districts $= 2\frac{2}{3}$

2. The variance within the district $= 3\frac{1}{6}$

Total $= 5\frac{5}{6}$

which is found to be equal to the variance per plot in the

state sample. Therefore, we know that the variance in the sample is equal to the sum of the variance between the districts and the variance within the districts. Therefore, if we know the variance of any two kinds, we can calculate the variance of the third type. Generally the variance within the districts is indirectly calculated by subtracting the variance between the districts from the variance in the sample.

Variety Approach (Ignore the presence of districts)

Let us now approach the problem from the varieties. For this purpose we shall first calculate the variance between the varieties by considering the average yield for each variety.

$\bar{y}$	$\bar{y}^2$
23	529
26	676
25	625
22	484
96	2314

(1) Mean $= 24;\quad \bar{\bar{y}}^2 = 576.$

$$(2)\ \text{Variance} = \frac{\Sigma\,\bar{y}^2}{N} - \bar{\bar{y}}^2$$

$$= \frac{2314}{4} - 576$$

$$= 578\frac{2}{4} - 576$$

$$= 2\frac{1}{2}$$

As in the case of districts, we can also calculate the variance within the varieties. For this purpose, the yield of each variety can be compared with the respective variety average yield. This can be calculated as follows :

	Variety I		Variety II		Variety III		Variety IV	
	y	y^2	y	y^2	y	y^2	y	y^2
	20	400	25	625	22	484	21	441
	23	529	26	676	25	625	22	484
	26	676	27	729	28	784	23	529
Total	69	1605	78	2030	75	1893	66	1454
Mean	$\frac{69}{3}$		$\frac{78}{3}$		$\frac{75}{3}$		$\frac{66}{3}$	
$\bar{y}$	23		26		25		22	
$\bar{y}^2$	529		676		625		484	

Variance

	Variety I	Variety II	Variety III	Variety IV
=	$\frac{1605}{3} - 529$	$\frac{2030}{3} - 676$	$\frac{1893}{3} - 625$	$\frac{1454}{3} - 484$
=	$535 - 529$	$676\frac{2}{3} - 676$	$631 - 625$	$484\frac{2}{3} - 484$
=	6	$\frac{2}{3}$	6	$\frac{2}{3}$

Total for all the 4 varieties $= 6 + \frac{2}{3} + 6 + \frac{2}{3} = 13\frac{1}{3}$

Average per variety $= 13\frac{1}{3} \div 4 = \frac{40}{3} \times \frac{1}{4} = \frac{10}{3} = 3\frac{1}{3}$

When we analyse the variance with reference to the varieties, we get two sets of variances as indicated below.

1. The variance between varieties $= 2\frac{1}{2}$
2. The variance within varieties $= 3\frac{1}{3}$

Total $= 5\frac{5}{6}$

This is equal to the variance per plot in the sample. As stated in the case of variance within the district, the variance within the varieties can also be calculated indirectly by subtracting the variance between the varieties from the total variance present.

(The following portion is not contemplated for the study. However this would be of interest to those who want to learn it.)

Simultaneous consideration of districts and varieties

Instead of considering the districts and varieties separately, let us now consider both these factors simultaneously. In this attempt we first consider the variance between the districts and the variance between the varieties. Normally, we expect that the sum total of these two variance should be equal to the total variance present in the sample.

Variance between the districts $= 2\frac{2}{3}$

Variance between the varieties $= 2\frac{1}{2}$

Total $= 5\frac{1}{6}$

This is not equal to the total variance present which is equal to $5\frac{5}{6}$. A shortage of $\frac{4}{6}$ or $\frac{2}{3}$ $\left(5\frac{5}{6} - 5\frac{1}{6}\right)$ is now noticed. We cannot give any valid reason for such difference. This may be due to certain factors which are acting beyond our control which we fail to control them in our experiment. The variance between the districts may be due to varieties and vice versa. Therefore, the only reason for this residual difference may be due to the combined effect of districts and varieties or the interaction between districts and varieties. Generally this is known as variance due to experimental error.

Therefore the total variance present in the population under study can be split up into three portions namely :

(1) The variance between districts $= 2\frac{2}{3}$

(2) The variance between varieties $= 2\frac{1}{2}$

(3) The variance due to the interaction between the districts and the varieties or experimental error $= \frac{2}{3}$

Total $= 5\frac{5}{6}$

Generally, the variance due to the experimental error is indirectly calculated by subtracting the total of variance between known factors such as districts and varieties in the problem from the totol variance present.

$$5\frac{5}{6} - \left(2\frac{2}{3} + 2\frac{1}{2}\right) = 5\frac{5}{6} - 5\frac{1}{6}$$

$$= \frac{4}{6} = \frac{2}{3}$$

DESIGN OF EXPERIMENTS

Problem

In agriculture, whether it is new varieties or cultivation practices or methods of treatments of seeds, a research worker has to conduct experiments mainly in the field. He has to try them in the field before he can compare their values. These objects of comparison in trials may be termed as treatments. The simple procedure of trying these 'treatments' each in different field or in different plot, is not sufficient enough to assess their relative value with reasonable confidence. If one conducts the treatments under the same conditions, one can find the inherent variation in the soil is quite considerable. Therefore, it may be sufficient to try the treatments on single plot side by side in the same field. A good idea of the

fertility variation can be obtained from the results of uniformity trials.

Uniformity trial

It consists of growing a particular crop in a field or piece of land with uniform treatment, by dividing the land into small units, and recording the produce of each of the units separately. We can find from the result of the yields that the fertility variation does not increase or decrease in any direction. On the other hand it may be distributed over the entire field in an erratic manner. However, there may be small homogeneous areas. Generally, the standard deviation of the yield gives an index of the inherent variability of the field.

Experimental Error

Apart from the uniformity ensured in respect of seed, sowing, cultivation practices, there may be other factors beyond the control of the expetimenter which may be responsible for the natural differences in fertility as reflected in the value of the standard deviation computed. Such variation from plot to plot due to uncontrolled factors is known as Experimental Error.

In order to allow for fluctuations due to experimental error, the research worker has to repeat the experiments many times. In the repetition of the experiments, if we find the difference once calculated persists consistantly we can accept the difference as real difference. Hence the difference is not due to fertility variation alone. When the treatments or experiments are repeated on a number of plots, the observed variation between the treatments may be partly due to the real difference of treatment and partly due to experimental error. The difference due to experimental error will have its influence on the results even if there is no real difference due to treatments. Hence it is necessary to compute the magnitude of the difference due to experimental error and compare it with that of the treatments so as to find out

whether there is any real difference in the effects of the treatments.

Replications

The repetition of the treatments under investigation is known as Replications. We cannot allow the effects of experimental error, directly to nullify our experiment. At the same time we cannot curb it also. Hence we have to average out its influence over the different treatments by means of replication. The procedure amounts to sampling.

Replication is necessary not only to stabilise the Mean but also for rigorous comparison of treatment effects. The fundamental reason is that only by replication we have means of estimating the experimental error.

Randomization

In order to have objective and effective comparison between treatments, it is also essential to have random allocation of the treatments to various plots instead of allocating them according to one's desire. Further, the statistical procedure adopted for comparison of the treatments will be valid only when the experiments are allocated randomnly among the plots.

By means of replication, the experimenter wants to average out the effects of environmental differences so as to give various treatments equal scope to show their real merit. This involves the question of arrangement of plots. By randomization we can ensure that the various treatments will be subject to equal environmental effect in the long run by repetition of the experiments.

Suppose we have four treatments A, B, C and D. A Randomization involves a systematic arrangements of plots. A common example of such systematic design is the chess board arrangement of plots such as

A B C D
D A B C
C D A B
B C D A

In this, all the treatments appear in each row as well as in each column. In this the influence of any fertility gradients along the sides of the rectangle will be eliminated. But it may be seen from the diagonal AA that the fertility is in favour of treatment A.

Even randomisation does not remove the difficulty in securing exactly equal environment for all treatments. Actual randominsation in any practical experiment may result in one of the very systematic arrangements. The merit of randomisation provides a rigorous basis for the test of significance of the difference between the treatments, compared with difference due to unequal environment.

Let us suppose that we have 20 plots of uniform size each and we are having two treatments, A and B. We want to try the treatments on a random basis.

Ten of these 20 plots may be allocated randomnly to one of the treatments. Suppose we have the following 10 random numbers for the treatment A, we can treat the plots corresponding to these random numbers with the treatment A and the remaining ten plots with treatment number B. We can test the significance of unit A and B from the results obtained.

15, 19, 13, 3, 6, 1, 8, 20, 10, 11.

Local control

Though the random allocation of treatments to plot gives an estimate of the treatment difference free from any systematic influence of the environment or bias and also provides a correct test of significance, it is not quite efficient. It is desirable to reduce the experimental error as far as possible and practical without disturbing the statistical randomness. This can be achieved by making use of the fact obtained earlier that adjacent areas are relatively more homogeneous than those widely separated. Therefore, instead of randomising the two treatments all over the field as done earlier, we can divide the 20 plots into 10 Blocks of 2 plots

each and allocate the treatments A and B randomnly within the Block. In this process the difference between A and B would be subject to the fertility variation within each Block alone. Generally, this variation would be less than that over the whole field.

This is due to the fact that the treatment difference is subject to variations between plot to plot only within Block. This variation is ganerally lower than plot to plot variation over the whole field.

Such arrangements in Blocks can be extended even if there are more than two treatments. Each group of contiguous plots forming a Block would contain as many plots as there are treatments. The treatments would be allocated among the plots in each block in a random manner. This arrangement is known as Randomised Block.

Experimental Design

Various forms of plot arrangements to suit the requirement of particular problems have been evolved and they are known as experimental design. The principle underlined in all these cases is same. It is to provide (by means of randomisation and replication) an unbiased comparison of treatments against their standard errors (Standard deviation of mean) and also to reduce the errors with the help of replication and local control.

Randomised Blocks

A sample application of the principles discussed before and one of the common uses in field trials is the design known as Randomised Block. The design is of wider applicability and several treatments can be tried together in the same way.

For this purpose the land on which the experiments to be tried out, should be divided into as many Blocks of same size and shape as there are replications. Each of the Blocks should thereafter be divided into as many plots of same size and shape as there are treatments. If there are 't' treatments and 'r'

replications, there should be 'r' Blocks and each Block should have 't' plots giving a total of 'tr' plots. The treatments are allocated randomly to the 't' plots in each Block with the help of random number. The 'tr' plot yields from the 'tr' plots would furnish the data for comparison of treatment.

Suppose we have 8 treatments and we want 6 replications, the following is one of the designs for the purpose.

I	II	III
4	6	3
6	2	8
3	3	2
8	7	4
7	5	7
5	8	1
1	1	5
2	4	6

IV	V	VI
2	4	2
3	7	7
7	6	6
6	2	5
5	8	3
1	3	1
8	1	8
4	5	4

The number given in various plots are nothing but random numbers and we have to allot the treatments corresponding to the random numbers.

Number of Replications

Greater care has to be taken to have the required number of replications so as to ensure efficiency.

Exercise

1. Explain analysis of variance.
2. Write short notes on:
   Variance between and variance within factors.
3. Analyse the variance into different components:

District	Variety		
	V_1	V_2	V_3
D_1	40	50	60
D_2	60	70	80
D_3	50	60	70

	Treatment				
	T_1	T_2	T_3	T_4	Total
V_1	25	30	45	40	140
V_2	35	45	25	35	140
V_3	30	40	40	50	160
V_4	50	45	50	55	200
Total	140	160	160	180	640

CHAPTER VII

TIME SERIES

We know that things which are capable of being represented in quantitative measures will not remain constant for ever, that is, the quantity may change from time to time. When we mean 'time' it refers to a period of time and that too to a long period and not a short period. The changes in the quantity may be due to many causes. As the causes for the variation are different, the types of change or variation are also different. What we are really interested is to find (1) the various types of variations noticed in the values of certain items and (2) the magnitude of the variations due to the different causes so that we can eliminate the effects of such causes and also forecast value of the items at a distant future. For this kind of study we require data for a long period of time or for a series of time and the study of such data due to various factors may be called Analysis of Time Series.

Movements in Time Series

Generally the variation can be broadly classified under four categories, namely Trend (T), Seasonal (S), Cyclic (C) and Irregular (I). Further it can be seen that these four types of variations may be combined either in an addition form or multiplication form to constitute the Time Series. In such cases the following formulae can be adopted.

Addition form of Time Series: $T+S+C+I$.

Multiplication form of TimeSeries: $T \times S \times C \times I$.

So, when we know the type of combination of these variations and the value of the variations, the value of the remaining variations can be obtained either by subtraction or by division as the case may be.

Secular Trend or Trend

In course of time, certain things may undergo changes in their value. We can consider the population of a country. Though there may not be any change in the area of the country, the population may go on increasing over a period of time (vide Table I). Due to increase in the size of the population, the demand for consumable articles may increase and consequently production of agricultural produce and industrial produce may increase. This type of changes noticed in the value by passage of time may be called 'Trend'. This change may be either upward or downward trend. If we consider the mortality among the people in a country, the death rate may decrease due to increased medical facilities available and also due to invention of new medicines and scientific advancement (vide Table 2). This shows that value will have a tendency to undergo change which may be due to many factors and this type of variations or movement may also be termed as Trend.

Table No. 1

Year	Population (in millions)
1957—58	1574
1958—59	1687
1959—60	1737
1960—61	1807
1961—62	1960
1962—63	2099
1963—64	2228
1964—65	2308
1965—66	2450

1966—67	2580
1967—68	2582
1968—69	2613
1969—70	2665

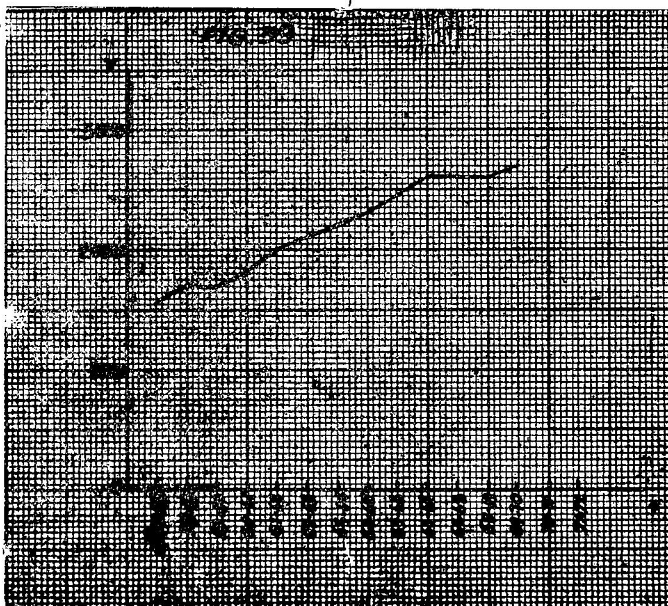


Fig. 7-1.

Table No. 2

Mortality Rate

Year	Rate per thousand people
1951	17·1
1952	16·0
1953	17·2
1954	14·0
1955	11·3
1956	13·6
1957	14·2
1958	13·1
1959	11·9
1960	12·1
1961	13·3
1962	11·3
1963	11·3
1964	10·8
1965	11·5
1966	11·0
1967	10·5
1968	8·7
1969	8·4
1970	8·2
1971	7·8

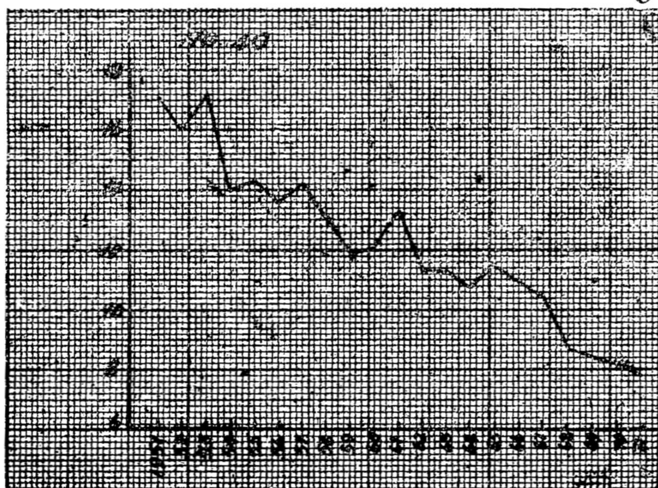


Fig 7-2.

Seasonal Movement or Periodic Movement

A periodic movement is one which recurs or repeats with some degree of regularity within a definite period. Generally movements noticed at definite intervals or in definite season in a year may be called a seasonal variation. Rainfall in a country is subject to seasonal variations. Similarly, the prices of agricultural commodities may decrease during harvest period and increase during the slack season. The sales of articles such as cloth etc., may increase during festival seasons especially in October to January. In the case of rain fall etc., nature is responsible for the seasonal variation while in the case of others, customs and festivals may account for

the seasonal variations. Banking transactions may be more on the day following a holiday. Sales will be heavy during the first week of every month due to disbursement of salary to workers in Government service and established firms.

Cyclical Movements

These movements are different from seasonal movements. While seasonsl movements recur at definite periodic intervals within a period of an year, the cyclical movements may repeat over a long period of time, say ten to fifteen years. However, they will not show a regular periodicity in their occurrence. Such kind of movements can be noticed in business circle. This may be due to the consequence of some sudden changes that may take place in some field, and naturally a chain of reactions may be noticed. Devaluation of currency in one country may also cause devaluation of currencies of other countries.

Irregular Movements

All movements other than those mentioned above can be termed as irregular movements since they do not exhibit a regular pattern in their occurrences. They may be due to an outbreak of some epidemic diseases or due to outbreak of war or due to some natural havoc such as cyclone, storms, earthquakes etc. Therefore such kind of variations are very difficult to foresee and assess.

Analysis of Time Series

The value of trend can be measured by any one of the following methods :—

(i) Free hand curve Method.

(ii) Moving Average Method.

(iii) Method of Least squares.

FREE HAND CURVE METHOD

Measurment of Secular Trend by Free Hand Curve Method

If we plot the data of a time series on a graph paper, with period marked on x-axis and value on the y-axis, we will have a curve representing the data. In the curve we can observe certain ups and downs in certain periods. Ignoring the presence of ups and downs if we draw another smooth free hand curve through as many points as possible, the new curve which smoothens the ups and downs will indicate the movement of the present trend in the data. This method is known as free hand curve method.

From this curve we can estimate the value at a particular time and this value may be called as 'estimated value' while the value given in the actual series will be called 'observed value'. Though there may be some difference or deviation between the observad value and the estimated value we can draw the curve in such a way that the overall difference is 0. This can be achieved by taking the average of the square of the differences by means of Least Square Method which will show that the average of a square of the deviation is minimum.

With the help of the Least Square Method we can also know the law of relationship and also fit a suitable curve by means of curve fitting method.

Example

Let us consider the following example :

Year	Production [in (000) tonnes]
1958	75
1959	78
1960	95
1961	112

1962	105
1963	115
1964	140
1965	128
1966	150
1967	165
1968	175
1969	170

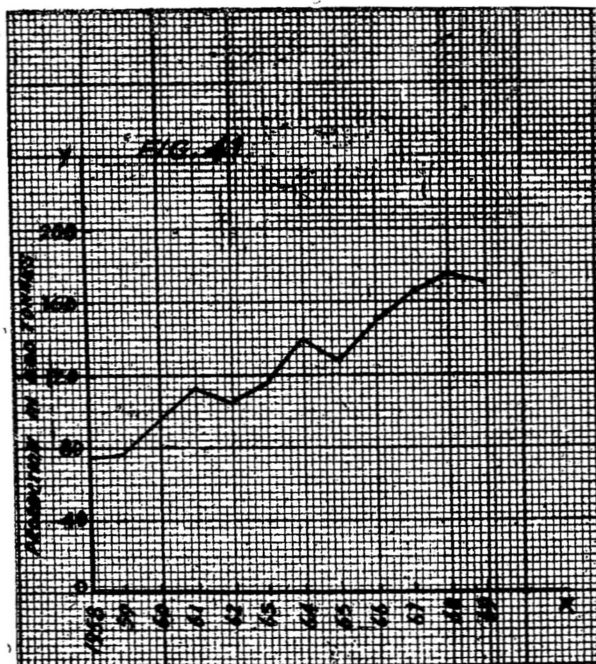


Fig. 7-3.

We can draw a trend line by free hand method for the above data. First we can plot the points and draw a curve for the data. The years can be marked on the x-axis and produc-

tion figures can be marked on the y-axis. After plotting the various points we can join them by means of straight line and this would be the curve for the data given.

·A straight line can be drawn such that the highest and the lowest points of the graph are approximately at equal distance from the trend line. The points on the trend line will represent the values.

After finding out a suitable trend for the given series as explanied earlier, we can determine the trend value for each year. The observed value for each year can thus be divided by the Trend value for the corresponding year and then multiplied by *100*. This will show the percentage of the original value in terms of Trend value. The trend can be eliminated.

This can be ensured that the vert'cal distance of the points, which are above the straight line, from the straight line are equal to the vertical distance of the points, which are below the straight line, from the straight line.

It may be noted that a monthly time series are typically the product of Secular Trend (T), Seasonal Variation (S), Cyclical Movements (C) and Irregular Movement (I) $(T \times S \times C \times I)$.

Merits of the Method

1. This is the simplest method of estimating trend.

2. The results can be quickly arrived at since no mathematical computation is involved and can easily be understood.

3. It can be used for all types of trend whether linear or non-linear.

4. This method eliminates the regular and irregular fluctuations. The line shows the basic tendency over a period of time.

5. An experienced person with sound knowledge of the economic history of the industry can handle this with ease and more accuracy.

Demerits

This is only a visual method of estimating the trend and hence cannot be used for correct prediction.

It is susceptible to the bias of the statistician since there are no specific rules. Hence various persons can draw different lines.

It requires special practice and good experience.

It is only approximation.

MOVING AVERAGE METHOD

The moving average method is a simple process to measure the trend. The meaning and method of its calculation can be examined with the help of an example

Example

The following data give the domestic consumption of power.

Year	Units consumed (in Million)
1970	247
1971	273
1972	276
1973	316
1974	395
1975	469
1976	501

For the construction of moving average, we can consider either a 3 years or 5 years or 7 years period as the case may be. Let us calculate a 3 years moving average. We should first take the first three years data and find out the average.

$\frac{247 + 273 + 276}{3} = \frac{796}{3} = 265.$ (This is the average for 1971.) Afterwards, omit 1970 and in its place add 1973 and calculate the average for 1971, 1972 and 1973.

$\frac{273 + 276 + 316}{3} = \frac{865}{3} = 288.$ (This is the average for 1972.) In this manner we can proceed till we exhaust the last figure.

$$\text{Average for } 1973 = \frac{276 + 316 + 395}{3} = \frac{987}{3} = 329.$$

$$\text{Average for } 1974 = \frac{316 + 395 + 469}{3} = \frac{1180}{3} = 393.$$

$$\text{Average for } 1975 = \frac{395 + 469 + 501}{3} = \frac{1365}{3} = 455.$$

Generally the moving averages are calculated from the moving total. This will be repeated as follows.

Year	Units consumed	3 years Moving total	3 years Moving Average
(1)	(2)	(3)	(4)
1970	247	—	—
1971	273	796	265
1972	276	865	288
1973	316	987	329

1974	395	1180	393
1975	469	1365	455
1976	501	—	—

Since we have taken three years moving average, one year in the beginning and one year at the end are not having any moving average. In case we take five years moving averages, 2 years in the beginning and 2 years at the end will not have any moving average.

Five years Moving Average

Year	Units consumed	5 years Moving total	5 years Moving Average
1970	247	—	—
1971	273	—	—
1972	276	1507	301
1973	316	1729	346
1974	395	1957	391
1975	469	—	—
1976	501	—	—

Centering of Moving Average

Sometimes, two years or four years moving averages are also calculated. In these cases the moving total and moving average will be entered between the successive pair of values. But this is inconvenient because the moving average does not exactly represent a year but represents a mid year. In order to get over this difficulty an adjustment is made so that the averages may coincide with a year. This type of adjustment is called centering the Moving Average. This can be done as follows:

Year	Production	4 years moving total	4 years moving average	2 years moving total of 4 years moving average	Central 4 years moving average
(1)	(2)	(3)	(4)	(5)	(6)
1960	80				
1961	85				
		343	86		
1962	90			178	89
		368	92		
1963	88			188	94
		383	96		
1964	105			198	99
		408	102		
1965	100			212	106
		440	110		
1966	115			225	113
		460	115		
1967	120			234	117
		478	119		
1968	125			245	123
		503	126		
1969	118				
1970	140				

A moving average can be defined as follows:

A moving average is an average of a fixed number of items in a time series which moves through the series by dropping the top item, of the previous averaged group and adding the next item below in each successive average.

Hence moving averages may be considered as an artificial time series in which each period's figure is replaced by the Mean of the value of that period and also those of a number of preceding and succeeding years.

Sometimes a Time Series may contain month-wise data. In such cases we can adopt twelve months moving averages. In the case of 12 months moving averages, the variations due to season can be smoothed out. The moving averages estimated on the basis of 12 months or 13 months will be written against the month in the middle of the year

In fact, the moving average will be considered as a rough estimate of the trend and cyclic movements because the movements due to season and to some extent the irregular movements are smoothed out. If the original data are divided by the moving average (T × C) we will have an estimate of the seasonal and irregular movements.

$$\text{Series} = T \times C \times S \times I$$

$$\text{Estimate} = \frac{T \times C \times S \times I}{T \times C} = S \times I$$

The selection of the period for calculating the moving average is an important problem. The main purpose is to get the trend value so that it is free from the effect of other types of fluctuations or subject to the minimum effect of the other fluctuations.

Merits of Moving Averages

1. It is a simple device to reduce fluctuations and obtain trend values with a fair degree of accuracy.

2. This method is not subject to personal bias as in the case of free hand method.

3. When the period of a cycle is taken as the moving average period, the cyclical variation can be eliminated.

Demerits

1. As the choice of the period requires great care, true trend value may not be obtained due to inappropriate method.

2. If the time series is a long one, the computation of moving average will be cumbersome.

3. As the moving average is based on Arithmetic Mean it is susceptible to extreme value.

4. We cannot have the trend value for some periods at both ends of the series. Hence it cannot be used for forecast.

METHOD OF LEAST SQUARES

The method of least squares is widely used and it is more popular to determine the trend values. With the help of Least Square method, we can also establish a mathematical law of relationship and also fit a suitable curve by means of curve fitting method. We have already studied about this method when we studied about Regression and Regression Lines. The method is same as fitting a straight line of the form $y = mx + c$ to the given data, where 'm' denotes the slope of the straight line with the x-axis. Generally it is expressed in terms of tangent value of the angle made by the straight line with x-axis and 'c' is the intercept made by the straight line on the y-axis.

Fitting a Linear or Straight line to the given data

The following is the production of a factory for the 1st nine years of its working.

Year (x) :	1	2	3	4	5	6	7	8	9
Production (000 tonnes) (y) :	25	30	28	35	42	40	47	49	55

Let the period be denoted by 'x' and the production be denoted by 'y'. Let us take the 5th year as the origin for our com-

putation so as to reduce the monotony of the computation. The other years can be expressed as a deviation from this central period and the details can be rearranged as follows. After converting the years in terms of the origin, the other columns namely x^2 and xy columns can be computed and given in the following table.

x	y	x^2	xy	
—4	25	16	—100	
—3	30	9	— 90	—281
—2	28	4	— 56	
—1	35	1	— 35	
0	42	0	——	
1	40	1	40	
2	47	4	94	501
3	49	9	147	
4	55	16	220	
0	351	60	220	

In this problem the following details are arrived at :

$n = 9$; $\Sigma x = 0$; $\Sigma y = 351$; $\Sigma x^2 = 60$; $\Sigma xy = 220$.

The equation is $y = mx + c$.

Where 'm' and 'c' are constants as far as these values are concerned,

we can calculate the value of 'c' by the following formula:

$$c = \frac{\Sigma y}{n}$$

$$= \frac{351}{9} = 39$$

We can also calculate the value of 'm' with the following formula:

$$m = \frac{\Sigma xy}{\Sigma x^2} = \frac{220}{60} = \frac{11}{3} = 3{\cdot}67$$

Hence the equation is, $y = 3{\cdot}67x + 39$

But when we give the equation, it is always necessary to give the origin of our estimation. In our case we have taken the fifth year as the origin of our estimation. So the equation will be $y = 3{\cdot}67\,x + 39$ with the fifth year as the origin.

The same problem can be given in the following form:

Year (x)	1961	1962	1963	1964	1965	1966	1967	1968	1969
Production (y)	25	30	28	35	42	40	47	49	55

We need not be alarmed by the years. Here we can take 1965 as the origin and the remaining years can be numbered as $-4, -3, -2, -1, 0, 1, 2, 3, 4$ as before. In this process also we get the same equation namely $y = 3{\cdot}67x + 39$. But the origin has to be given as 1965. Therefore the equation is $y = 3.67\,x + 39$ with 1965 as the origin.

Estimation of Trend Value

The object of fitting a straight line to the given data is to find out the trend values. Now let us find out the trend values.

The method is simple. In the equation we have to substitute the value of 'x' for which the corresponding 'y' value is to be estimated.

$$\text{Equation is } y = 3.67x + 39$$

For 1961, $x = -4$

$\therefore \quad y = 3.67 \times (-4) + 39 \quad = \quad 24.32$

1962 : $\quad x = -3$

$y = 3.67 \times (-3) + 39 \quad = \quad 27.99$

1963 : $\quad x = -2$

$y = 3.67 \times (-2) + 39 \quad = \quad 31.66$

1964 : $x = -1$

$y = 3.67 \times (-1) + 39 \quad = \quad 35.33$

1965 : $x = 0$

$y = 3.67 \times (0) + 39 \quad = \quad 39.00$

1966 : $x = 1$

$y = 3.67 \times 1 + 39 \quad = \quad 42.67$

1967 : $x = 2$

$y = 3.67 \times 2 + 39 \quad = \quad 46.34$

1968 : $x = 3$

$y = 3.67 \times 3 + 39 \quad = \quad 50.01$

1969 : $x = 4$

$y = 3.67 \times 4 + 39 \quad = \quad 53.68$

It may be noted that the trend equation need not be used each time to compute the trend value. It is enough if we use the equation for the first value. Afterwards if we add the value of 'm' to the preceding value of 'y' we can get the succeeding value.

Let us compare the 2 values for y.

x	y_o	y_c	$y_o - y_c = d$	d^2
1961	25	24·32	0·68	0·4624
1962	30	27·99	2·01	4·0401
1963	28	31·66	—3·66	13·3956
1964	35	35·33	—0·33	0·1089
1965	42	39·00	3·00	9·0000
1966	40	42·67	—2·67	7·1289
1967	47	46·34	0·66	0·4356
1968	49	50·01	—1·01	1·0201
1969	55	53·68	1·32	1·7424
	351	351·00	0	37·3340

We find that though there are differences between the individual observed (y_o) and the computed trend values (y_c) the total difference is found to be *'0'*. Consequently, the average difference will also be *'0'*. Besides these, the sum of the squares of deviation and consequently the Mean Square Deviation will be the minimum or least. Hence this method is called 'Least Square Method'.

In the above example the number of years or in other words the number of period is 9 which is an odd figure. If the number of years or number of periods is even such as 8 or 10 some difficulty will be encountered in fixing the mid point of 'n' since there will be two mid years instead of one, such as 4 or 5. In such cases, instead of taking a particular year as the origin (mid year) we have to select the mid year between the two mid years. The computation will also undergo changes. Let us consider the following example :

Year x	y
1961	25
1962	30
1963	28
1964	35
1965	42
1966	40
1967	47
1968	49
1969	55
1970	59

As there are 10 years we have to select 1965 and 1966 as the mid periods since they are the central years. But we cannot have two years as the origin. Hence we should select the mid year between 1965 and 1966 as the origin and denote it as '0'.

Consequently, the different years will be converted into the following values in terms of deviation from the origin.

	x	y	x^2	xy
1961	—4·5	25	20·25	—112·5
1962	—3·5	30	12·25	—105·0
1963	—2·5	28	6·25	— 70·0
1964	—1·5	35	2·25	— 52·5
1965	—0·5	42	0·25	— 21·0
1966	0·5	40	0·25	20·0
1967	1·5	47	2·25	70·5
1968	2·5	49	6·25	122·5
1969	3·5	55	12·25	192·5
1970	4·5	59	20·25	265·5
		410	82·50	310·0

After converting the periods in terms of deviation from the origin we can calculate the values of x^2 and xy present there in the form of a table as given above.

$$n = 10; \quad \Sigma y = 410; \quad \Sigma x^2 = 82.50; \quad \Sigma x y = 310$$

As before we can fit a straight line of the form $y = mx + c$. We can use the following formula for finding out the value of m and c.

$$c = \frac{\Sigma y}{n} = \frac{410}{10} = 41$$

$$m = \frac{\Sigma xy}{\Sigma x^2} = \frac{310}{82.5} = 3.76$$

Therefore the equation is $y = 3.76\,x + 41$

with the middle of 1965—66 as the origin.

In this process the same amount of difficulties are experienced in the computation of x^2 and xy values because of the values 0.5, 1.5, 2.5 etc. when compared with the previous example. Even this difficulty can be overcome by the following substitution :

	x
1961	9
1962	7
1963	5
1964	3
1965	1
1966	1
1967	3
1968	5
1969	7
1970	9

Instead of taking 1 year period as the unit we can take six months or ½ year as 1 period and consequently the x values will undergo changes as given above. Further computation will be as follows :

	x	y	x^2	xy	
1961	−9	25	81	−225	
1962	−7	30	49	−210	
1963	−5	28	25	−140	− 722
1964	−3	35	9	−105	
1965	−1	42	1	− 42	
1966	1	40	1	40	
1967	3	47	9	141	
1968	5	49	25	245	+ 1342
1969	7	55	49	385	
1970	9	59	81	531	
		410	330	620	

$$e = \frac{\Sigma y}{n} = \frac{410}{10} = 41$$

$$m = \frac{\Sigma xy}{\Sigma x^2} = \frac{620}{330} = 1.88$$

Equation is $y = 1.88x + 41$ (with 1965-66 as the origin) of period six months.

Let us compare the two equations calculated with 1 year as the period of computation and ½ year as the period of computation. In both the cases we have taken the mid year of 1964-65 as the origin.

(1) $y = 3.76\,x + 41$ with 1964-65 as the origin.

(2) $y = 1.88\,x + 41$ with 1964-65 as the origin.

[Period (½) half year.]

We find practically no difference between the two equations. Since we have taken one year or 12 months as the period in the first equation $m = 3.76$ which is equal to twice the value of 'm' (1.88) in the second equation when the period is six months or ½ year. Therefore the second method is preferable to the first because of the easy computation.

Note: It is always necessary to mention the origin and the period whenever the equation to the trend line is estimated.

Merits of the Least Square Method

1. We can get the trend values for all the years.

2. We can also compute the value for any period not in the series. It means we can forecast the value for future years.

3. The sum of the deviation of the trend values from the actual values given is 0 and the average of deviation is 0 and hence it gives the best estimates of the trend values.

4. It is free from personal bias and it is most objective method since it is based on mathematical law.

Demerits

It may take time for computation.

SEASONAL VARIATIONS

In this, we shall study about the seasonal variation of the Time Series. When the data given are annual values, there will not be any seasonal variations since seasonal variations appear at weekly, monthly or quarterly intervals. The factors responsible for seasonal changes are climate or weather or festivals or customs. Seasonal variations can be estimated by the following methods :

1. Method of Simple Average.
2. Method of Moving Average.

Method of Simple Average

This is the simplest method. Suppose we are given monthly values or data for various years, first we must arrange the data in a systematic manner. In the first column we should enter the names of the 12 months. In the remaining columns we should enter the years.

Month	1961	1962	1963	1964	Total	Average
January						
February						
March						
April						
May						
June						
July						
August						
September						
October						
November						
December						
Total						

After recording the monthly figures we should find out the total of each month and enter in the last column against the respective mouths. In this manner we will have 12 totals for 12 months. Each monthly total should be divided by the number of years and the monthly average should be arrived at and it should be entered in the last column against the month. Afterwards the total of the monthly average should be calculated. This total should be divided by 12 to arrive at the general average. Each of the monthly averages should be divided by the general average or overall average and the result obtained should be multiplied by *100*. In other words, each monthly average should be expressed as a percentage of the general average. The percentage values thus arrived are called the seasonal index explaining the seasonal variation.

$$\text{Seasonal Index of a month} = \frac{\text{Monthly average}}{\text{General average}} \times 100$$

Let us consider the following example:

Months/Year	1966	1967	1968	1969	1970	Total	Average
January	342	355	182	255	911	2045	409
February	298	417	190	285	655	1845	369
March	259	343	197	325	471	1595	319
April	293	322	193	314	478	1600	320
May	352	316	170	348	444	1630	326
June	426	392	158	434	465	1875	375
July	497	305	263	510	460	2035	407
August	547	286	225	486	496	2040	408
September	604	295	236	493	522	2150	430
October	731	301	295	562	576	2465	493
November	642	260	266	675	557	2400	480
December	588	198	260	804	530	2380	476
Total							4812
Average							401

Month	Monthly Average	Seasonal Index
January	409	102
February	369	92
March	319	80
April	320	80
May	326	81
June	375	93
July	407	101
August	408	102
September	430	107
October	493	123
November	480	120
December	476	119
Total	4812	1200
Average	401	100

We shall consider an example where the data are given on a quarterly basis instead of monthly basis.

Year	First Quarter	Second Quarter	Third Quarter	Fourth Quarter
1970	42	45	47	48
1971	40	42	45	47
1972	38	37	40	41
1973	40	38	36	39
1974	45	48	42	40
Total	205	210	210	215
Average	41	42	42	43

We should first calculate the average for each quarter. Afterwards we should calculate the general average as follows:

Then each quarterly average should be divided by the general average and expressed as a percentage which will be the seasonal index.

		Seasonal Index
First Quarter	41	$\frac{41}{42} \times 100 = 97.7$
Second Quarter	42	$\frac{42}{42} \times 100 = 100.0$
Third Quarter	42	$\frac{42}{42} \times 100 = 100.0$
Fourth Quarter	43	$\frac{43}{42} \times 100 = 102.4$
Total	168	
Average	42	

The same procedure should be followed if we are given weekly or daily details. In the same manner we can calculate seasonal index for weekly or daily data.

Moving Average Method

In the method of simple averages, it is indirectly assumed that the effects of trend and cyclical variations on the time series are insignificant. So the original monthly values in the series are taken as the estimates of seasonal variations and averaged. In taking the average, we eliminate the random fluctuation from the time series and hence we get the seasonal variation.

In the case of Moving Average method, we do not assume the effects of trend and cyclical variations as insignificant. Hence we first calculate the two parts namely trend and cyclical variation by computing the moving average from time series. The period of moving average is taken as 1 year. Afterwards, the original monthly figures are divided by the moving averages of the concerned month and the original figure is expressed in terms of percentage of the moving average. The trend and cyclical variations are eliminated from the time series.

After converting the monthly figures into percentages of the moving averages, the percentages of each month are added separately and average is calculated for each month. Afterwards an average for all the months is calculated. Then each monthly average is expressed as percentage of the overall average by dividing each monthly average by overall average and then multiplying by *100*. The resultant figures are taken as the seasonal variations.

Let us assume the multiplication model and enumerate the various steps involved in the calculation of seasonal variation by the method of moving average.

1. If monthly data are given, we should first calculate 12 months moving average.

2. The moving average value should be first centred. The centred moving average values give trend and cyclical variations. It is free from seasonal variations since seasonal variations recur at regular intervals of one year or leas than one year. Since we have taken 12 months or 1 year moving average, the seasonal variations are eliminated.

3. Each monthly value in the original series should be divided by the corresponding centred moving averages and expressed as a percentage. These percentage values are the estimates of the seasonal variation for each month.

4. These percentages are to be rearranged so as to enable us to arrive at the total and average for each of the 12 months.

5. We should find the total value for each month and from this we should find the average for each month.

6. From the monthly average for all the 12 months we should calculate the general or overall average for a month.

7. Each monthly average should then be divided by the overall average and expressed as a percentage of the overall average and thus the seasonal index is calculated for each month.

$$\text{Seasonal Index} = \frac{\text{Monthly Average}}{\text{General Average}} \times 100$$

Monthly Prices of Chillies — in Rs.

Virudhunagar

Month	1965—66	1966—67	1967—68	1968—69	1969—70	1970—71
April	192·80	293·00	322·00	192·75	314·50	477·50
May	192·50	352·00	314·95	169·05	347·65	443.65
June	176.75	426·25	392·00	159·00	433·50	465·00
July	198·00	496·50	304·50	163·13	511·00	462·00
August	198·00	546·75	286·25	225·20	485·60	496·25
September	198·00	604·00	295·20	236·25	492·75	520·00
October	198 00	730·50	301·25	295·00	562·00	574·00
November	383·00	641·50	261·25	266·00	675·00	557·25
December	352·00	588·00	198·60	260·00	800·75	529·50
January	341·75	355·00	182·50	255·00	912·50	523·00
February	298·25	418·75	190·25	285·33	655·00	422·33
March	258·75	342·60	197·20	325·25	470·75	380·00

Calculation of Moving Average

Month	Price	12 months moving total	Total of two 12 months moving total	Centring of the 12 months moving average	Percentage of prices in terms of the moving average
(1)	(2)	(3)	(4)	(5)	(6)
1965					
April	192·80				
May	192·50				
June	176·75				
July	198·00				
Aug.	198·00				
Sep.	198·00				
		2987·80			
Oct.	198·00		6075·80	253·16	78·21
		3088·00			
Nov.	383·00		6335·50	263·98	145·09
		3247·50			
Dec.	352·00		6744·50	281·25	125·20
		3497·00			
1966					
Jan.	341·75		7292·50	303·85	112·47
		3795·50			
Feb.	298·25		7939·75	330·82	90·10
		4144·25			
March	258·75		8694·50	362·27	71·40
		4550·25			
April	293·00		9633·00	401·38	73·00
		5082·75			
May	352·00		10424·00	434·33	81·04
		5341·25			

(1)	(2)	(3)	(4)	(5)	(6)
1966					
June	426·25		10918·50	454·94	93·69
		5577·25			
July	496·50		11167·75	465·32	106·70
		5590·50			
August	546·75		11301·50	470·90	116·11
		5711·00			
Sept.	604·00		11505·85	479·41	125·99
		5794·85			
October	730·50		11618·70	484·11	150·90
		5823·85			
Nov.	641·50		11610·65	483·77	132·60
		5786·80			
Dec.	588·00		11539·35	480·81	122·29
		5752·55			
1967					
Jan.	355·00		11313·10	471·38	75·31
		5560·55			
Feb.	418·75		10860·60	452·53	92·54
		5300·05			
March	342·60		10291·30	428·80	79·90
		4991·25			
April	322·00		9553·25	398·05	80·89
		4562·00			
May	314·95		8743·75	364·32	86·45
		4181·75			

(1)	(2)	(3)	(4)	(5)	(6)
1967					
June	392·00		7974·10	332·25	117·98
		3792·35			
July	304·50		7412·20	308·84	98·59
		3619·85			
August	286·25		7011·20	292·13	97·99
		3391·35			
Sept.	295·20		6637·30	276·55	106·74
		3245·95			
Oct.	301·25		6362·65	265·11	113·60
		3116·70			
Nov.	261·25		6087·50	253·65	103·00
		2970·80			
Dec.	198·60		5708·60	237·86	83·49
		2737·80			
1968					
Jan.	182·50		5334·23	222·26	82·11
		2596·43			
Feb.	190·25		5131·81	213·83	88·97
		2535·38			
March	197·20		5011·81	208·83	94·43
		2476·43			
April	192·75		4946·61	206·11	93·50
		2470·18			
May	169·05		4945·11	206·05	82·04
		2474·93			

(1)	(2)	(3)	(4)	(5)	(6)
1968					
June	159·00		5011·26	208·80	76·15
		2536·33			
July	163·13		5145·16	214·38	76·09
		2608·83			
August	225·20		5312·74	221·36	101·73
		2703·91			
Sept.	236·25		5535·87	230·65	102·42
		2831·96			
Oct.	295·00		5787·67	241·07	122·37
		2953·71			
Nov.	266·00		6086·02	253·58	104·90
		3132·31			
Dec.	260·00		6539·12	272·46	95·43
		3406·81			
1969					
Jan.	255·00		7161·49	298·40	85·46
		3754·68			
Feb.	285·33		7769·76	324·00	88·14
		4015·08			
March	325·25		8286·66	345·28	94·20
		4271·58			
April	314·50		8810·16	367·09	85·67
		4538·58			
May	347·65		9486·16	395·26	87·95
		4947·58			

(1)	(2)	(3)	(4)	(5)	(6)
1969					
June	433·50		10435·91	434·83	99·69
		5488·33			
July	511·00		11634·16	484·76	105·41
		6145·83			
Aug.	485·60		12661·33	527·56	92·05
		6515·50			
Sept.	492·75		13176·50	549·02	89·75
		6661·00			
October	562 00		13485·00	561·88	100·02
		6824·00			
Nov.	675·00		13743·80	572·66	117·87
		6919·80			
Dec.	800·71		13871·10	577·46	138·55
		6951·30			
1970					
Jan.	912·50		13853·60	577·23	158·08
		6902·30			
Feb.	655·00		13812·25	575·64	113·79
		6912·95			
March	470·75		13853·15	577·21	81·56
		6940·20			
April	477·50		13892·40	578·85	82·49
		6952·20			
May	443·45		13786·65	574·44	72·20
		6834·45			

(1)	(2)	(3)	(4)	(5)	(6)
1970					
June	465·00		13397·65	558·24	83·30
		6563·20			
July	462·00		12737·40	530·73	87·05
		6174·20			
Aug.	496·25		12115·73	504·82	98·30
		5941·53			
Sep.	520 00		11792·31	491·74	105·83
		5850·78			
Oct.	574·00				
Nov.	557·25				
Dec.	529·50				
1971					
Jan.	523·50				
Feb.	422·33				
March	380·00				
April					
May					
June					

Calculation of Seasonal Index Numbers

	Oct.	Nov.	Dec.	Jan.	Feb.	March	April	May	June	July	Aug.	Sep.
1. (1965-66)	78·2	145·1	125·2	112·5	90·1	71·4	73·0	81·0	93·7	106·7	116·1	126·0
2. (1966-67)	150·9	132·6	122·3	75·3	92·5	79·9	80·9	86·5	118·0	98·6	98·0	106·7
3. (1967-68)	113·6	103·0	83·5	82·1	89·0	94·4	93·5	82·0	76·2	76·1	101·7	102·4
4. (1968-69)	122·4	104·9	95·4	85·5	88·1	94·2	85·7	88·0	99·7	105·4	92·1	89·8
5. (1969-70)	100·0	117·9	138·6	158·1	113·8	81·6	82·5	77·2	83·3	87·1	98·3	105·8
Total	565·1	603·5	565·0	513·5	473·5	421·5	415·6	414·7	470·9	473·9	506·2	530·7
Average	113·0	120·7	113·0	102·7	94·7	84·3	83·1	82·9	94·2	94·8	101·2	106·1

Month	Average	Seasonal Index
October	113·0	$\frac{113 \cdot 0 \times 100}{99 \cdot 2} = 113 \cdot 9$
November	120·7	$\frac{120 \cdot 7 \times 100}{99 \cdot 2} = 121 \cdot 6$
December	113·0	$\frac{113 \cdot 0 \times 100}{99 \cdot 2} = 113 \cdot 9$
January	102·7	$\frac{102 \cdot 7 \times 100}{99 \cdot 2} = 103 \cdot 5$
February	94·7	$\frac{94 \cdot 7 \times 100}{99 \cdot 2} = 95 \cdot 4$
March	84·3	$\frac{84 \cdot 3 \times 100}{99 \cdot 2} = 85 \cdot 0$
April	83·1	$\frac{83 \cdot 1 \times 100}{99 \cdot 2} = 83 \cdot 8$
May	82·9	$\frac{82 \cdot 9 \times 100}{99 \cdot 2} = 83 \cdot 6$
June	94·2	$\frac{94 \cdot 2 \times 100}{99 \cdot 2} = 94 \cdot 9$
July	94·8	$\frac{94 \cdot 8 \times 100}{99 \cdot 2} = 95 \cdot 5$
August	101·2	$\frac{101 \cdot 2 \times 100}{92 \cdot 2} = 102 \cdot 0$
September	105·1	$\frac{106 \cdot 1 \times 100}{99 \cdot 2} = 106 \cdot 9$
Total	1190·7	1200·0
General Average	99·2	100·0

Additive Model

We have considered the Multiplicative Model. If it is additive model, the following procedure may be followed.

1. After arriving at the moving average, we should subtract the moving average from the corresponding monthly values and find out the deviation.

2. The deviation for the 12 months should be arranged so as to enable us to arrive at the total as well as the monthly average deviation. These average deviations will serve as the seasonal variations.

CYCLICAL MOVEMENTS

Cyclical variations are regular as in the case of seasonal variations. While the period of seasonal variations are less than one year, the period of cyclical variation is more than one year and usually it varies from 5 to 10 years. Seasonal variations recur at regular intervals. But cyclical variations do not take place at regular intervals.

Secular trends represent movements over a long period of time. But cyclical movements represent movements over a period of 5 to 10 years.

Secular trends represent continuous movements in the same direction either increasing or decreasing. But cyclical movements represent both increasing and decreasing trend.

Cyclical movements are very common in business and hence known as business cycles. We know that there are ups and downs in business. There are well defined periods in business cycles. They are (1) Depression (2) Recovery (3) Prosperity or Boon (4) Recession or Decline.

The period from one depression to the next depression or from one recovery to next recovery or from one prosperity to next prosperity or from one recession to another recession is called a cycle.

Measurement of Cyclical Movements

Once we compute the values of the two components namely Trend (T) and Seasonal (S) variations, we can get cyclical variations.

Multiplicative Model

(1) Trend values are first calculated by the method of Moving Averages or by the method of Least Squares. These values are denoted by the letter 'T'.

(2) Seasonal indices are calculated by the method of simple average or moving average method and these indices are denoted by 'S'. (It may be noted that a monthly time series are typically the product of secular trends (T), seasonal variations (S), cyclic movements (C) and irregular movements. $Y = T \times S \times C \times I$.)

Any series which contains only annual figures will not contain Seasonal Variation since Seasonal Variation can be seen only in the monthly figures. Similarly, annual figures will be free from irregular movements also.

(3) Hence if we divide the original values by the product of T and S, we will get the product of $C \times I$. The product of Trend and Seasonal Index is $Y = T \times S \times C \times I$.

$$\therefore \frac{Y}{T \times S} = C \times I$$

(4) We can calculate the moving average from the value of $C \times I$. These moving averages will give us the cyclical component of the given time series, since in calculating the moving average we are removing the irregular movement (I) from $C \times I$.

However, this can be smoothened by the use of short term moving average. Irregular movements may be short term duration, say, monthly and only occasionally it may be of a long

time duration. Hence a two months or a three months moving average of the percentages representing cyclical and irregular movements can be adopted to eliminate irregular movements and arrive at the cyclic movements also.

Irregular Movements

This is the residual movement which cannot be explained satisfactorily in terms of the other components. These are due to irregular or accidental fluctuations like strikes, lockouts, floods, wars etc. There is no regular period or time for their occurrence. Because of this irregular character it is very difficult to isolate them. There is no method to isolate irregular fluctuations from the original data. After removing trend, seasonal variations and cyclical variations from the given data the residual can be taken as the effects of irregular movements.

Exercise

1. What is a Time Series? Mention its components and also explain the nature of their combination.

2. Explain Time Series analysis. Indicate its importance in business.

3. Write short notes on:

   (1) Trend (2) Seasonal variation. (3) Cyclical movements (4) Moving averages (5) Least square methods.

4. Describe the method for determining the trend.

5. What are the common methods for eliminating seasonal variation from Time Series.

6. Calculate three years moving average for the following time series and plot it with the original figures on the same graph.

Year	Output
1946	150
1947	140
1948	150
1949	210
1950	260
1951	320
1952	350

7. Fit a straight line for the following data.

Year	Quantity
1960	70
1961	75
1962	80
1963	85
1964	90
1965	95
1966	100

8. Find out the Seasonal Index from the following data.

Season	1960	1961	1962	1963	1964
1st Quarter	40	42	41	45	44
2nd Quarter	35	37	35	36	38
3rd Quarter	38	39	38	36	38
4th Quarter	40	38	42	41	42

CHAPTER VIII

DIFFERENT TYPES OF SAMPLE SURVEYS

The purpose of taking a sample is to get an estimate of a desired characteristic with an error as low as possible for a fixed cost or with as small a cost as possible for a fixed margin of error in the estimate. Several types of sampling procedures have been developed. This will help either in reducing the sampling error or in reducing the cost or both. We shall study about some of the important types. In this, we shall confine our study to the selection of sample units only without estimation procedure.

It should also be noted that in all different sampling procedures the selection of the sample or drawing of the sample for collecting the information is being done with the help of the random numbers to avoid bias. Therefore, in all types of surveys only random samples are selected.

Selection of Random Samples and use of Random Numbers

We have already studied about the use of random numbers in the selection of random samples. The best method commonly used at present is the use of Random Numbers. There are different sets of published Random Numbers namely Tippett's Random Numbers, Fisher and Yates Random Numbers.

The process of drawing a random sample by making use of random numbers is to identify each sampling unit in the population with a number starting from 1 to N with the help of these random numbers. We should select random number either equal to or less than N. After selecting the required numbers we should select the units having the serial numbers corresponding to the random numbers selected.

Random Sample with Replacement

In this procedure we choose 'n' units from the population of 'N' units. The unit which is once selected is replaced again before the next sample unit is selected. The procedure is similar to the one described under sample without replacement except for the difference that repeated units are accepted as many times as they occur during the process of selection. In a sample of 'n' units with replacement, there may be 'n' or less than 'n' distinct (different) units.

Example 1 — Procedure

Suppose we have a taluk consisting of 5 firkas containing 98 revenue villages, and we want to select 5 revenue villages for a socio-economic survey. First we should give serial numbers to all the 98 revenue villages commencing from 1 to 98. Since the total number, i.e. 98, is a two digit figure we should consult 2 digit random numbers for the selection of sample.

Let us consult col (2) in 2 digit random numbers.

The random numbers selected are 51, 97, 79, 69, 60. Since the first 5 random numbers are less than 98 we can take all these 5 random numbers. The revenue villages to be selected are those with serial numbers 51, 97, 79, 69 and 60.

Example 2

In the above example we have suggested that all the villages in all the revenue firkas have to be given the serial numbers. Sometimes it may not be necessary. Suppose we know that the total number of revenue villages in each firka is as given below:

Firka No.	Total No. of villages in the firka.	Cumulative Total No. of villages.	First Sl.No.	Last Sl. No.
1	20	20	1	20
2	15	35	21	35
3	18	53	36	53
4	25	78	54	78
5	20	98	79	98

As per the random numbers we have to select the following villages.

51, 97, 79, 69 and 60.

It may be seen that the following villages are in the firka noted against each.

Revenue Village		Firka
51		3
60,	69	4
79,	97	5

Since the villages selected are from the firkas 3, 4, 5 it in enough if we give serial numbers to the villages only in these firkas. In this process we save time.

Stratified Sampling

In certain cases the units in the population may be hetro-geneous in character and in such cases the population will be divided into different groups, with or without equal number of units. However, units of each group will be more or less homogeneous within the group. Each group is called a stratum and hence it is called a stratified sampling. After stratification, the required number of samples are taken from each stratum considering each stratum as a seperate universe. The number of samples to be selected from each stratum may differ depending upon the size of the stratum or the number of units in the stratum. Hence the chance for the selection of one unit within a particular stratum may not be same as the chance of another unit in another stratum. However, when we consider the whole universe, the chance of selection of one unit in any stratum will be same as the unit in any other stratum. This may be due to the fact that the chance of one unit in a stratum depends first upon the chance of that stratum in the universe. When these chances are taken together simultaneously the chance of the unit in each stratum will be uniformly the same. The stratification of the universe may be done on the basis of some other characteristics which is closely related to the character under study. If we want to divide a district into

different parts for the purpose of study of area under a particular crop, the district can be divided into many groups depending upon the entire area rather than the area of the particular crop, since the extent of a particular crop in an area may depend upon the extent of the entire area itself.

Example

Let us suppose that there are 600 workers in a factory and their weekly wage ranges from Rs, 6 to Rs. 100. How can we select the sample?

Procedure

We can divide the workers into 5 categories based upon the wages. Let it be as follows :

Wages Range	No. of workers	Size of a sample no. of workers
Less than Rs. 10	180	9
Rs. 10 — Rs. 25	120	6
Rs. 26 — Rs. 50	120	6
Rs. 51 — Rs. 80	100	5
Rs. 81 and above	80	4
Total	600	30

Suppose we want to estimate the income on the basis of 5% sample, we have to select 9 workers from Group I, 6 workers from Groups II and III, 5 workers from Group IV and 4 workers from Group V.

Strata	Column construc-ted	Random No. selected	Sl. No. of the worker to be selected
(1)	(2)	(3)	(4)
I Stratum			
Total No. of workers 180. (180 × 5 = 900)	2	807	807 ÷ 180; R = 87
		186	186 ÷ 180; R = 6
		410	410 ÷ 180; R = 50
		345	345 ÷ 180; R = 165

(1)	(2)	(3)	(4)
Random Numbers to be consulted 1 to 900.		626	626 ÷ 180; R = 86
(Total samples		340	340 ÷ 180; R = 160
to be selected 9.)		883	883 ÷ 180; R = 163
		569	569 ÷ 180; R = 29
		341	341 ÷ 180; R = 161
II			
Total No. of			
workers 120.	2	094	= 94
(120 × 8 = 960)		322	322 ÷ 120; R = 82
Random Numbers to be consulted		252	252 ÷ 120; R = 12
1 to 960.		047	= 47
(Total samples to be		469	469 ÷ 120; R = 109
selected 6.)		632	632 ÷ 120; R = 32
III			
Total No. of workers 120.	3	270	270 ÷ 120; R = 30
(120 × 8 = 960) Random Numbers to		608	608 ÷ 120; R = 8
be consulted 1 to 960.		099	= 99
		226	226 ÷ 120; R = 106
(Total samples to		225	225 ÷ 120; R = 105
be selected 6.)		928	928 ÷ 120; R = 88
IV			
Total No. of workers 100.	3	273	273 ÷ 100; R = 73
(100 × 9 = 900)		858	858 ÷ 100; R = 58

Randoms Numbers to be consulted 1-900.		221	221 ÷ 100; R = 21
(Total samples to		479	479 ÷ 100; R = 79
be selected 5.)		243	243 ÷ 100; R = 43
V			
Total No. of	3	212	212 ÷ 80; R = 52
workers 80.			
(80 × 12 = 960)		384	384 ÷ 80; R = 64
Random Numbers to be consulted 1-960.		233	233 ÷ 80; R = 73
(Total samples to be selected 4.)		569	569 ÷ 80; R = 9

Note : The number given in the last column is the remainder obtained by dividing the random number by the total number of workers in each stratum.

Systematic Sampling

In this process all the units in the samples selected for the survey will not constitute random samples. Only the first unit to be selected will be a sample selected with the help of random numbers and the subsequent units will not be random samples. After selecting the first unit with the help of random numbers, the subsequent units will be selected at constant interval from one another depending upon the total number of units to be selected for the survey a which in turn depend upon the proportion of the samples to the population or in other words sampling fraction.

As explained earlier, we should first give serial numbers to all the units in the population.

Suppose there are 150 fields in a village and we want to select 5% sample fields for our survey,

total number of fields $= 150$

5% sample $= \frac{150 \times 5}{100} = 7\frac{1}{2}$

or 8 fields approximately.

Since it is a 5% sample it amounts to 1 in 20. Therefore the inter-space between one sample unit and another unit should be 20.

As we have to select eight fields so as to constitute 5%, let us first select a random number which is either equal to or less than 8. Since 8 is a one digit number we should consult one digit random numbers. Let us consult the fourth column in the one digit random numbers. The first number is 6. Since it is less than 8, we must select 6. The first field should be the 6th field. The subsequent fields should be,

1) $6 = 6$

2) $6 + 20 = 26$

3) $26 + 20 = 46$

4) $46 + 20 = 66$

5) $66 + 20 = 86$

6) $86 + 20 = 106$

7) $106 + 20 = 126$

8) $126 + 20 = 146$

The serial numbers of the fields to be selected are 6, 26, 46, 66, 86, 106, 126 and 146.

Cluster Sampling

In this survey, the samples to be selected will consist of different clusters each one of which may consist of more than one unit or a few equal number of units. Only one unit in each cluster will be selected with the help of random numbers

and the remaining units in the cluster will not be seleted on the basis of random numbers.

At first, required number of sample units will be selected with the help of random numbers. After selecting a unit, a few units in and around the selected unit will be selected to form a cluster and the number of units in each cluster may be the same. This is generally resorted to if the size of the population is vast and extensive in character. In the case of surveys for estimating the yield of coconut and arecanut trees, generally cluster sampling is being adopted. Suppose we want 3 clusters of each 5 trees for our survey and the random numbers selected are 11, 27, 43, the serial number of trees selected are 1st cluster 9, 10, 11, 12, 13; 2nd cluster 25, 26, 27, 28, 29; 3rd cluster 41, 42, 43, 44 and 45.

Line Sampling

Sometimes the population may consist of a number o parallel rows of lines as trees in a garden or houses in a colony. In such cases, a particular unit or row may be selected with the help of random numbers and a required number of trees or houses may be selected in that selected row. In this method, we need not first enumerate all the trees in the garden or houses in the colony for preparing the frame. Instead, it is enough if we prepare the list of rows and confine the preparation of the list of houses or trees only to that particular row selected.

Multi-staged sampling

In multi-staged sampling the universe is considered as consisting of a number of first stage units each of which is made up of a number of second stage units and so on. The sampling process is carried out at different stages and hence called multi-staged sampling. The type of sampling is less accurate. However it has its own advantages. The construction of the final stage sampling has to be done only for that group at the penultimate stage, which is going to be selected for the final selection.

At present, crop cutting experiments are conducted only by this method. Each district is composed of many taluks and each taluk is considered as a stratum. Each taluk consists of many villages and each village consists of many survey numbers. Each survey number may consist of many sub-divisions. Each sub-division may consist of many fields and each field may consist of many plots of our required size for the experiment. After selecting the taluk by random, a village in the selected taluk is selected and a survey number in that village and a sub-division in that survey number and a field in that sub-division and finally a plot in the field is selected. This involves sampling at different stages or multi-stages. Hence it is a multi-staged sampling.

Sampling with varying probabilities

In the case of random sampling the principle underlined is that each and every unit in the population should be given equal chance or probability of being selected. But in certain surveys this principle is not strictly followed due to consideration of certain other factors connected with the study.

In some cases the difference between the values among the individual units may be very wide and application of equal probabilities may not give a good estimate of the characteristic of the population. In such cases the individual units have to be given importance or probability according to their value.

In a few circumstances the value of characteristics of our study may not be readily available. However they may be correlated with some other characteristic of the population. For example, the area under a particular crop, say, paddy in a village may depend upon the irrigated area in each village. In such circumstances the villages to be selected for our survey can depend upon the irrigated area in the village. This can be done as follows:

First, we should give serial numbers to all the units in the population. Against each unit we should record the

value of the related characteristic (irrigated area). Afterwards, we should find the cumulative values of this characteristic and enter the cumulative value against each unit. We shall see the example given below :

Sl. No. of village.	Irrigated area	Cumulative area	Probability
	('000' acres)		
1.	75	75	75/860
2.	55	130	55/860
3.	48	178	48/860
4.	57	235	57/860
5.	99	334	99/860
6.	125	459	125/860
7.	73	532	73/860
8.	76	608	76/860
9.	95	703	95/860
10.	157	860	157/860
	860		

We should find the total value of this characteristic of the population. In this case it is 860 (N). Suppose we have to select 3 (n) units, we should consider 860 as the highest number and select 3 random numbers from the 3 digit random numbers. Suppose we consult 5th column of 3 digit random number, we would get random 029, 265 and 689. Therefore, we should select the units corresponding to the serial numbers 29, 265 and 689. But the serial numbers are only imaginary numbers since there are no such serial numbers already. But we can consult the cumulative values. From the cumulative value we know that 29 is contained in the first village, 265 in the 5th village and 689 in the 9th village. Therefore, we should select the 1st, 5th and 9th villages. In this process, there is a possibility of selecting one village more than once with replacement. The disadvantage in this method is the

difficulty in cumulating the values, when the items and the values are large.

Exercise

1. Write an essay on the different types of sample surveys.

2. Write short notes on:

    1) Stratified sample
    2) Systematic sample
    3) Cluster sampling
    4) Multi-stage sampling

3. There are 325 households in a village. Select a suitable sampling procedure and select a 5% sample households for a socio-economic survey.

4. There are 238 rows of trees in a garden each consisting of 15 trees. Suggest a suitable sampling procedure for estimating the yield by means of 2% sample trees.

5. There are 185 houses in a village. Select a sample of 10 houses by means of systematic sample.

6. There are 6 taluks in a district and the population in each taluk is given against each. Select 2 taluks with probability proportional to the size of the population.

Taluk	Population
1	150000
2	250000
3	350000
4	450000
5	550000
6	750000

ONE-DIGIT RANDOM NUMBERS

Columns

(1)	(2)	(3)	(4)	(5)
3	3	2	6	1
2	7	0	7	3
1	3	5	5	3
5	7	1	2	1
0	6	1	8	4
8	7	3	5	2
2	1	7	6	3
1	2	8	6	7
1	5	5	1	0
9	0	5	2	8
0	6	7	6	5
2	0	1	4	8
3	2	9	8	9
8	0	2	2	0
5	4	4	2	0

TWO - DIGIT RANDOM NUMBERS

Columns

(1)	(2)	(3)	(4)	(5)
51	51	00	83	63
68	97	87	64	81
30	79	20	69	22
81	69	40	23	72
90	60	73	96	53

(1)	(2)	(3)	(4)	(5)
46	15	38	26	61
99	05	48	67	26
98	35	55	03	36
11	53	44	10	13
06	71	95	06	79
83	45	19	90	70
49	90	65	97	38
39	84	51	67	11
16	17	17	95	70
13	74	63	52	52

THREE - DIGIT RANDOM NUMBERS

Columns

(1)	(2)	(3)	(4)	(5)
642	807	270	546	029
790	186	608	897	265
435	410	099	205	689
218	345	226	433	905
263	626	225	267	531
296	340	928	403	526
835	883	273	307	700
058	569	858	422	469
452	341	221	191	226
757	094	479	348	407
149	322	243	302	047
639	252	212	801	325
648	047	384	924	748
573	469	233	958	782
879	632	569	615	352

SECOND YEAR — 2nd PAPER

PRACTICALS

A. DIAGRAMMATIC REPRESENTATION

Diagrammatic representation — Bar Charts — Component bars — Adjacent bars — Percentage bar diagrams —Pie diagrams.

1. The following data relate to the monthly expenditure of 2 families. Represent the data by drawing suitable Bar diagrams, Bar charts, Component bars, Adjacent bars, Percentage bar diagram, Pie charts and compare the pattern of expenditure of the two families.

ITEM	FAMILY A	FAMILY B
Income (p.m.)	**Rs. 500**	**Rs. 400**
	Rs.	Rs.
Food	210	160
Clothing	80	80
House rent	100	60
Education	30	40
Fuel & Lighting	40	20
Miscellaneous	40	40

2. Draw a suitable diagram to represent the following data. The figures indicate the investment in the Second Five Year Plan.

Sectors	Rs. in Crores
Agriculture	568
Irrigation and power	913
Industry and Mining	890
Transport	1385
Social services	945
Miscellaneous	99
Total	4800

3. From the data given below, construct a chart showing the shift in the distribution of population in a country between urban and rural areas. Give a brief comment on the chart.

Year	Population in million	
	Urban	Rural
1900	14	36
1910	22	41
1920	30	46
1930	42	49
1940	54	51
1950	69	54
1960	75	56

4. Marks obtained by 40 students are given below. Construct a frequency table choosing appropriate class interval. Draw a Histogram, Frequency polygen and Frequency curve-Ogive.

56, 24, 89, 42, 56, 72, 91, 96, 43, 32,
19, 62, 75, 66, 54, 48, 52, 82, 36, 62,
41, 37, 85, 72, 66, 54, 34, 41, 27, 39,
68, 53, 74, 81, 29, 61, 49, 36, 86, 81.

5. The following table gives the marks of 100 students in Statistics. Draw the Ogive curve and find the Median.

Marks	Frequency
70 – 80	5
60 – 70	6
50 – 60	20
40 – 50	31
30 – 40	22
20 – 30	9
10 – 20	7
	100

6. The following are the weights of 50 bundles (in kg). Prepare a suitable frequency table and a cumulative frequency table and draw less than Ogive curve.

42, 74, 40, 60, 82, 115, 41, 61, 75, 63, 68,

53, 110, 76, 84, 50, 67, 65, 78, 77, 56, 95,

69, 104, 80, 79, 79, 54, 73, 59, 81, 100, 66,

49, 77, 90, 84, 76, 42, 64, 69, 70, 80, 72,

50, 79, 52, 103, 96, 51.

7. The following table gives the height of certain plant in a group. Draw the Ogive and calculate the Median.

Heights (in cms.)	Frequency
58	3
60	10
62	27
64	40
66	26
68	20
70	9
72	8
74	7

8. Form a frequency distribution by taking suitable class-interval for the following data giving the weight of 50 students in a class room.

67, 34, 36, 48, 49, 31, 61, 34, 43, 45, 38, 32, 27,

61, 29, 47, 36, 50, 46, 30, 46, 32, 30, 33, 45, 49,

48, 41, 53, 36, 37, 47, 47, 30, 46, 50, 28, 35, 35,

38, 36, 46, 43, 34, 62, 69, 50, 28, 44, 43.

9. Draw a Lorenz curve for the following data.

(i) Average Income Rs.	Number of workers	(ii) Average Expenditure Rs.	Number of workers
45	5	40	4
58	6	50	7
65	8	60	9
75	9	70	6
80	2	75	4

B. MEASURES OF CENTRAL TENDENCIES

1. The table below gives the charges for grinding of flour of selected companies. Calculate the average net cost of grinding a barrel of flour. Calculate the median cost of grinding.

Cost of grinding flour per barrel Rs.	No. of companies operating
4.40 – 4.79	14
4.80 – 5.19	15
5.20 – 5.59	35
5.60 – 5.99	19
6.00 – 6.39	10
6.40 – 6.79	4
6.80 – 7.19	2
7.20 – 7.59	1
	100

2. Compute some of the measures of Central tendencies from the following data.

Class interval Rs.	Frequency
155 – 157	4
158 – 160	8
161 – 163	26
164 – 166	53
167 – 169	89
170 – 172	62
173 – 175	48
176 – 178	14
179 – 181	6

3. Calculate the Arithmetic Mean of the following frequency distribution of 700 working class families.

Income per week Rs.	No. of families
110 – 115	60
115 – 120	120
120 – 125	210
125 – 130	201
130 – 135	70
135 – 140	25
140 – 145	11
145 – 150	3
	700

4. The following table gives the height of certain variety of plants. Find the Arithmatic Mean, Median and Mode.

Height cms.	No. of plants
30 – 39	15
40 – 49	46
50 – 59	75
60 – 69	53
70 – 79	40
80 – 89	18
90 – 99	3

5. Calculate the Mean, Mode for the following frequency table.

Size of the item	Frequency
Below 4	3
4 – 5	8
5 – 6	28
6 – 7	59
7 – 8	66
8 – 9	27
9 – 10	6
above 10	3
	200

6. Determine the Arithmatic Mean and the Mode of the following frequency distribution.

Class interval (cms)	Frequency
16 – 17	3
17 – 18	13
18 – 19	23
19 – 20	31
20 – 21	18
21 – 22	9
22 – 23	2
23 – 24	1

7. Find the Median aud Mode of the following frequeney distribution.

Class interval (Rs.)	Frequency
0 – 5	10
5 – 10	12
10 – 15	17
15 – 20	20
20 – 25	20
25 – 30	18
30 – 35	11
35 – 40	10

8. Calculate the Mean and Mode of the following distribution.

Class interval (kg)	Frequency
10 – 20	5
20 – 30	9
30 – 40	13
40 – 50	21
50 – 60	20
60 – 70	15
70 – 80	8
80 – 90	3

9. Calculate the Mean, Median, Mode of the following distribution.

Class interval (Rs.)	Frequency
20 – 25	50
25 – 30	70
30 – 35	100
35 – 40	180
40 – 45	150
45 – 50	120
50 – 55	70
55 – 60	60

10. The mean salary paid to 100 employees is found to be Rs.180. It was discovered afterwards that the salary of two persons were wrongly entered as Rs. 297 and 165 instead of the correct salary, Rs. 197 and 185. Find the correct mean.

11. Calculate the quartiles Q_1 and Q_3 from the following data. Also calculate from the following other measures, first and third quintiles.

Calculate D_4, D_7, D_9, P_{20}, P_{70} and P_{90}

15, 35, 10, 47, 25, 52, 37, 42, 48.

12. Calculate the following measures of Central tendencies from the following data.

Value Rs.	Frequency
20	4
27	7
35	9
38	15
45	8
50	6

(i) Q_1 (ii) Q_3 (iii) First quintiles (iv) 4th quintiles (v) D_8 (vi) D_5 (vii) P_{20} (viii) P_{40} (ix) P_{70}

13. Calculate the following measures from the table below.

(i) Q_1, Q_3 (ii) 1st and 3rd quintiles (iii) D_5, D_7 (iv) P_{20}, P_{50}, P_{70}.

Class (cms.)	Frequency
0 – 5	4
5 – 10	7
10 – 15	9
15 – 20	15
20 – 25	8
25 – 30	7

14. Calculate the Geometric Mean for the following data.

(i) 250, 489, 353, 757, 982

(ii) 983, 1250, 456, 7951, 2845

(iii) 450, 987, 1215, 395, 285

15. Calculate the Geometric Mean for the following data.

	Value	Frequency
(i)	25	4
	38	7
	49	4
	35	2
(ii)	290	2
	457	3
	625	5
(iii)	195	3
	252	4
	172	8

16. Calculate the Hormonic Mean for the following data.

(i) 25, 86, 45, 42, 50

(ii) 49, 25, 40, 50, 12

(iii) 5, 7, 9, 10, 15

17. Calculate the Harmonic Mean for the following data.

(i)

x	f
20	4
30	5
40	2
50	4

(ii)

x	f
25	2
35	4
55	7
45	6
65	1

C. MEASURES OF DISPERSION

1. Find the Standard Deviation for the following frequency distribution.

Height in cms.	**No. of children**
59 – 61	3
61 – 63	12
63 – 65	54
65 – 67	111
67 – 69	128
69 – 71	85
71 – 73	30
73 – 75	6
75 – 77	1

2. Compute the Standard Deviation of the distance travelled by 260 farmers to buy certain daily necessities.

Km travelled	No. of farmers
1	19
3	52
5	70
7	39
9	24
13	21
15	14
17	12
19	9

3. Calculate the Semi - Inter Quartile range for the following distribution of wages.

Weekly wages (Rs.)	Frequency (No. of workers)
40 – 43	4
43 – 46	15
46 – 49	27
49 – 52	36
52 – 55	24
55 – 58	18
58 – 61	9
61 – 64	7

4. Calculate the co-efficient of variation for the following distribution of wages.

Wages per week in Rs. Mid value	Frequency (No. of workers)
38	27
44	72
50	135
56	170
62	285
68	175
74	96
80	28
86	12

5. Find the Mean Deviation for the following frequency distribution with (1) Mean (2) Median as the origin.

Length Mid value (cm)	Frequency (No. of units)
4.0	2
4.2	7
4.4	10
4.6	35
4.8	50
5.0	90
5.2	52
5.4	26
5.6	12
5.8	9
6.0	7

6. Calculate the Mean Deviation and Standard Deviation for the following data.

Length of calls per minute	No. of calls
0 – 1	12
1 – 2	30
2 – 3	21
3 – 4	16
4 – 5	11
5 – 6	5
6 – 7	2
7 – 8	2
8 – 9	1

7. Compute the Standard Deviation and Quartile Deviation for the following table.

Age (years)	No. of persons
20 – 25	33
25 – 30	112
30 – 35	152
35 – 40	154
40 – 45	136
45 – 50	118
50 – 55	96
55 – 60	74
60 – 65	54
65 – 70	37
70 – 75	34

8. Compute Quartile Deviation for the following frequency table

Size (cm)	Frequency
4 – 8	6
8 – 12	10
12 – 16	18
16 – 20	30
20 – 24	15
24 – 28	12
28 – 32	10
32 – 36	6
36 – 40	3

9. Find the Standard Deviation and the co-efficient of variation for the following data.

Class interval (kg)	Frequency
0 – 10	5
10 – 20	10
20 – 30	20
30 – 40	40
40 – 50	30
50 – 60	20
60 – 70	10
70 – 80	5

10. Calculate the Standard Deviation for the following table.

Class interval (years)	Frequency
25 – 34	4
35 – 44	20
45 – 54	38
55 – 64	24
65 – 74	10
75 – 84	4

11. The runs scored by 2 players in 10 innings are given below. Who is more consistent ?

A	25	65	45	0	50	100	35	80	10	90
B	40	55	50	35	50	65	45	60	40	60

D. FITTING A STRAIGHT LINE

1. Obtain a line of best fit for the following data.

Year	Consumption in tonnes
1920	27
1921	29
1922	28
1923	31
1924	30
1925	32
1926	36
1927	37
1928	38
1929	40
1930	42

2. Fit a straight line for the following data and with its help estimate the value for 1951.

Year	Population (in million)
1881	23
1891	31
1901	39
1911	50
1921	63
1931	76
1941	92

E. CORRELATION CO-EFFICIENT

1. Calculate the co-efficient of correlation for the following.

Year	Value of the raw cotton exported (Rs. in crores)	Value of cotton goods imported (Rs. in crores)
1915 – 16	42	56
1917 – 18	44	49
1919 – 20	58	53
1920 – 21	55	58
1923 – 24	89	65
1929 – 30	96	76
1931 – 32	66	58

2. The following details are given.

Mean $\bar{x} = 65$; Mean $\bar{y} = 67$

Standard Deviation $\sigma_x = 3.5$

Standard Deviation $\sigma_y = 2.5$

Correlation co-efficient $r = 0.8$

(i) Write down 2 regression lines.

(ii) Obtain the best estimate of x when $y = 70$

3. The regression lines of y on x and x on y are given below.

$y = 0.80\,x + 25$

$x = 0.45\,y + 30$

Find the correlation co-efficient between x and y.

4. The regression lines of y on x and x on y are given below.

$y = 0.9\,x + 2.3$

$x = 0.4\,y + 0.86$

Find the co-efficient of correlation between them.

5. The table below gives the age of 12 pairs of husband and wife. Calculate the correlation co-efficient.

Age of husband	Age of wife
25	18
22	15
28	20
26	17
35	22
20	14
22	16
40	21
20	15
18	14
19	15
25	23

6. The following table gives the height of fathers and sons. Compute the co-efficient of correlation.

Father's Height (in cm.)	Son's height (in cm.)
167	165
168	166
164	167
167	168
172	168
170	169
170	171
169	172
173	173

7. In a correlation table, the regression lines are given by

$$5x = 6y + 20$$

$$100y = 768x - 3608$$

Find the correlation co-efficient between x and y.

8. Find the regression lines from the following data.

$\bar{x} = 125$, $\sigma_x = 15$

$r = 0.55$

$\bar{y} = 80$, $\sigma_y = 9$

9. Find the co-efficient of correlation between the variation of x and y from the following table.

x	y
57	113
59	117
62	126
63	126
64	130
65	129
55	111
58	116
57	112

10. Find the co-efficient correlation between the variation of x and y from the table given below:

x	y
50	102
51	107
52	106
53	108
54	113
55	117
56	127
58	134
61	136

11. The prices of 2 commodities appear to be fluctuating together. The prices of these commodities over a period of time are given. Calculate the measure of relationship between the prices.

Year	Price of commodity A Rs.	Price of commodity B Rs.
1960	85	64
1961	65	71
1962	77	85
1963	88	60
1964	99	71
1965	102	85
1966	87	69
1967	71	70

F. RANK CORRELATION

1. The ranks awarded by two professors for 10 students are as follows. Calculate the rank correlation co-efficient among the marks awarded.

Students	Prof. A	Prof. B
1	1	3
2	6	5
3	5	8
4	10	4
5	3	7
6	2	10
7	4	2
8	9	1
9	7	6
10	8	9

2. The ranks obtained by 10 students in 2 papers are as follows. Calculate the rank correlation co-efficient.

1st Paper	2nd Paper
3	6
5	4
8	9
4	8
7	1
10	2
2	3
1	10
6	5
9	7

3. The marks obtained by 10 students in 2 papers are as follows. Calculate the rank correlation co-efficient.

1st Paper	2nd Paper
45	71
70	68
41	35
49	32
50	48
25	43
40	58
62	57
65	70
48	65

G. ANALYSIS OF VARIANCE

The following are the results of yield (kg) obtained in 12 experimental plots on four varieties of paddy in three districts. Analyse the variance in the following manner.

1. (a) Variance between varieties.
   (b) Variance within varieties.
2. (a) Variance between districts.
   (b) Variance within districts.

District	Varieties			
	1	2	3	4
1	20	25	22	21
2	23	26	25	22
3	26	27	28	23

3. Find out the variance, with the help of the following data (weight in kg).
   (a) between and within varieties.
   (b) between and within treatments.

Variety	Treatment			
	1	2	3	4
1	20	25	32	35
2	35	30	25	22
3	25	30	35	26
4	40	35	20	21

H. TESTS OF SIGNIFICANCE

1. A die is thrown 50 times. The probability of a success is 1/3. The number of success in an experiment is 20. Find out whether the die is biased.

2. The average height of students in a class is 150 cm and their Standard Deviation is 15 cm. The mean height of the students in a school of 400 strength is 155 cm. Does this indicate any significant difference ?

3. A test was conducted in a large group of students and the Standard Deviation of the score was found to be 25. The

test was conducted among boys and girls and their average scores are as follows:

	Boys	Girls
Number	25	36
Average score	120	140

Find out whether there is significant difference between the score of boys and girls.

4. In a countrywide investigation, the incidence of a particular disease is found to be 2%. In a college of 500 strength, 15 students are found to be affected by this disease and in another college of 1,500 strength, 10 students are affected by this disease. Find out whether any significant difference exists.

5. An examination of the writings of a particular author revealed that 5% of the words used are of foreign language. In a passage containing 6,000 words of the same author 50 words are found to be from foreign language. Does this indicate any significant difference ?

6. A renowned tyre company has advertised that their tyres will have an average running of 16,000 km without any repair, with a Standard Deviation of 1,500 km. When a lot of 100 tyres was purchased, their average running was found to be 15,500 km. Can we say this lot belongs to the same company ?

7. A tube light company has advertised that their lights will have an average burning of 8,000 hours with a Standard Deviation of 500 hrs. When we purchase 50 lights we find that their life is 8,500 hrs. Can these 50 lights be the products of the same company ?

8. When a coin is tossed 100 times, head occurs on 65 occasions. Can this be a good coin ?

I. ASSOCIATION OF ATTRIBUTES

1. Calculate the co-efficient of association between food habit and eye sight on the basis of the data of 1,620 persons among the age group 60 - 70.

Eye Sight	Food-Habit	
	Vegetarian	Non-Vegetarian
Defective	200	107
Normal	813	500

2. Find out the association between inoculation and immunity from the attack.

	Affected	Non-affected
Inoculated	10	80
Not-inoculated	20	40

3. From the following data, find out whether the Attributes A and B are independent.

$(A) = 100 \qquad (B) = 120$

$(AB) = 40 \qquad (N) = 300$

4. Given the following ultimate class frequencies, find the frequencies of the positive and negative classes and the whole number of observation N.

$(AB) = 200; \; (A\beta) = 100; \; (B\alpha) = 160; \; (\alpha\beta) = 80.$

5. Show whether A and B are independent, positively associated and negatively associated in the following cases.

(i) $N = 1{,}000; \; (A) = 470; \; (B) = 620; \; (AB) = 320.$

(ii) $(AB) = 512; \; (B\alpha) = 96; \; (A\beta) = 288; \; (\alpha\beta) = 255$

(iii) $(A) = 245; \; (AB) = 147; \; (\alpha) = 285; \; (B\alpha) = 190.$

6. Find out the co-efficient of association between the types of colleges training the students and the success in teaching from the following table:

	Passed	Failed	Total
Teachers' College	40	60	100
University	55	45	100
	95	105	200

J. INDEX NUMBERS

1. Calculate the index number of prices from the following data.

Commodity	1935		1945	
	Price Rs.	Quantity	Price Rs.	Quantity
A	4	50	10	40
B	3	10	9	2
C	2	5	4	2

2. Calculate Fisher's Ideal Index Number from the following table.

Commodity	Price in Rs.		Quantity	
	1955	1956	1955	1956
Rice	20.00	15.00	1 quintal	1.25 quintal
Salt	4.00	4.75	10 litres	8 litres
Cloth	10.50	12.50	20 metres	18 metres
House Rent	10.00	12.00	per month	per month

3. Construct a suitable Price Index Number for the year 1970 from the following data taking 1960 as the base year.

Commodity	1960		1970	
	Price Rs.	Quantity Quintal	Price Rs.	Quantity Quintal
A	4	1	10	2
B	1	10	4	25
C	20	2	90	3
D	10	5	15	20

4. Construct Fisher's Ideal Index Number for the following data.

Commodity	Base Year Price Rs.	Base Year Quantity (kg)	Current Year Price Rs.	Current Year Quantity (kg)
A	5	1	10	3
B	3	10	6	25
C	20	3	60	4
D	10	6	15	20

5. The following table gives the data of production in tonnes and price per ton of four groups with 1960-61 as the base. Calculate the Fisher's Ideal Index Number for the price of 1969-70.

Commodity	Production 1960-61	Production 1969-70	Price Rs. 1960-61	Price Rs. 1969-70
A	250	300	150	130
B	100	120	120	200
C	20	30	600	1000
D	10	20	200	300

K. TIME SERIES

1. The following data give the index numbers of export value in India. Smooth the data by fitting a linear trend by the method of moving average taking 5 years period.

Year	Index No.	Year	Index No.
1938 – 39	100.0	1946 – 47	284.9
1939 – 40	119.8	1947 – 48	372.2
1940 – 41	130.3	1948 – 49	421.4
1941 – 42	155.9	1949 – 50	435.7
1942 – 43	184 6	1950 – 51	482.9
1943 – 44	227.4	1951 – 52	711.7
1944 – 45	244.3	1952 – 53	500.0
1945 – 46	240.8	1953 – 54	461.0

2. The number of books borrowed from a public library on six working days are given below for a period of 6 weeks. Compute the seasonal (daily) variation in the series.

S. No. of week	Number of books borrowed					
	Mon.	Tues.	Wed.	Thurs.	Fri.	Satur.
I	25	43	49	46	51	62
II	18	34	52	49	53	70
III	12	25	48	51	62	66
IV	19	22	49	61	71	72
V	21	32	43	53	61	72
VI	13	30	29	46	50	60

3. Fit a straight line trend for the following data on the demand of motor fuel.

Year	Quantity—'000 litres
1946	61
1947	66
1948	72
1949	76
1950	82
1951	90
1952	96
1953	100
1954	108
1955	110
1956	114

4. Fit a straight line of the form $y = mx + c$ for the following data.

Year	Production in lakhs
1961	8
1962	12
1963	15
1964	18
1965	20
1966	23
1967	27
1968	30

5. Compute 3 years moving average and determine the long trend.

Year	Price (Rs.)
1920	97
1921	109
1922	108
1923	112
1924	113
1925	110
1926	115
1927	116
1928	118
1929	119
1930	120
1931	122
1932	124

6. The sales of a firm during 7 consecutive years are as follows. Fit a linear trend and give the estimate of sales for the 8th year.

32, 45, 36, 78, 94, 112, 136 (in lakhs of Rs.)

L. VITAL STATISTICS

1. Calculate the crude death rate in the following cases.

	Mid year population	Deaths
(i)	1870	17
(ii)	1925	19
(iii)	200050	1890
(iv)	19005	165
(v)	17000	170

2. Calculate the crude birth rate—

	Mid year population	Births
(i)	19250	195
(ii)	20050	350
(iii)	19500	320
(iv)	20000	450
(v)	25000	420

3. The following is the age distribution of the population. Calculate the specified death rate for the age groups 40 – 50, 50 - 60 and 60 – 70 for men and women separately.

Age	Men	Deaths	Women	Deaths
0 – 10	18900	287	19000	100
10 – 20	17000	295	18000	250
20 – 30	20000	300	20500	350
30 – 40	40000	356	39000	400
40 – 50	29000	258	28000	350
50 – 60	19000	250	18500	300
60 – 70	10000	325	9500	350
Above 70	9000	350	8300	400

M. LIFE TABLE

1. Construct a mortality table starting from 20 years upto 30 years for the following data.

x	lx
20	10000
21	9870
22	9740
23	9600
24	9470
25	9330
26	9190
27	9050
28	8900
29	8740
30	8580

(a) Find the probability that one person aged 20 lives for 10 more years.

(b) Find the probability that one person aged 25 dies in his 30th year.

2. Fill the blanks in the following table.

Age	lx	dx	px	gx	Lx	Tx
30	762230	—	—	—	—	—
31	758580	—	—	—	—	—

N. SAMPLE SURVEYS

1. There are 235 houses in a village. Select, with the help of random number,

(a) a sample of 5% houses.

(b) a sample of 10% houses.

2. There are 15 rows, each containing 25 houses in a newly developed colony. Select a row and select a sample of 5% houses in the row.

3. There are 25 rows of coconut trees in a garden. Select a cluster of 5 trees for yield estimation by means of simple sample.

4. There are 95 trees scattered over the garden. Select a sample of 10 trees by means of systematic sampling.

5. There are 250 survey numbers in a village. Select five clusters of each three survey numbers for detailed survey.

6. There are 10 taluks in a district and the number of villages in each taluk are as follows. Select a sample of three villages by means of multi-stage sample.

Taluk	No. of villages
1	42
2	57
3	65
4	28
5	40
6	53
7	93
8	70
9	25
10	77

7. Select a sample of one taluk from the above with probability proportional to the size of the taluk in respect of villages.

O. PROBABILITY

1. 5 balls are drawn from a bag containing 4 red and 6 black balls. What is the probability for the following occurences?

   (i) 2 balls are red and 3 black balls.

   (ii) 3 balls are red and 2 black balls.

   (iii) 4 balls are red and 1 black ball.

   (iv) 1 ball is red and 4 black balls.

   (v) All the 5 balls are black.

2. A bag contains 3 white balls, 5 black balls and 6 red balls A ball is drawn at random. What is the probability that it is either red or white ?

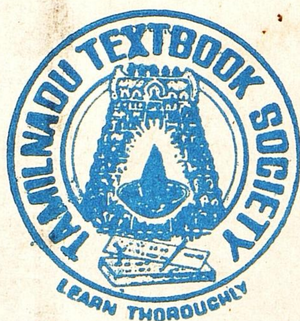
3. A throw has been made with 2 dice. What is the probability that the sum of the numbers thrown will be 10 ?

4. Two numbers are chosen at random from the set of numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12. What is the probability that the sum is equal to 8 ?

5. The sum of 2 positive integers is 10. What is the probability that the product does not exceed 20 ?

6. 5 persons are selected from a group containing 10 men, 5 women and 6 children. Find the probability that exactly 3 of them are women.

7. Find the probability of drawing 4 white balls and 2 black balls without replacement from a bag containing 1 red, 4 black and 6 white balls.


TAMILNADU TEXTBOOK SOCIETY
LEARN THOROUGHLY